# DM and DW on Clickstream Mining : A case study

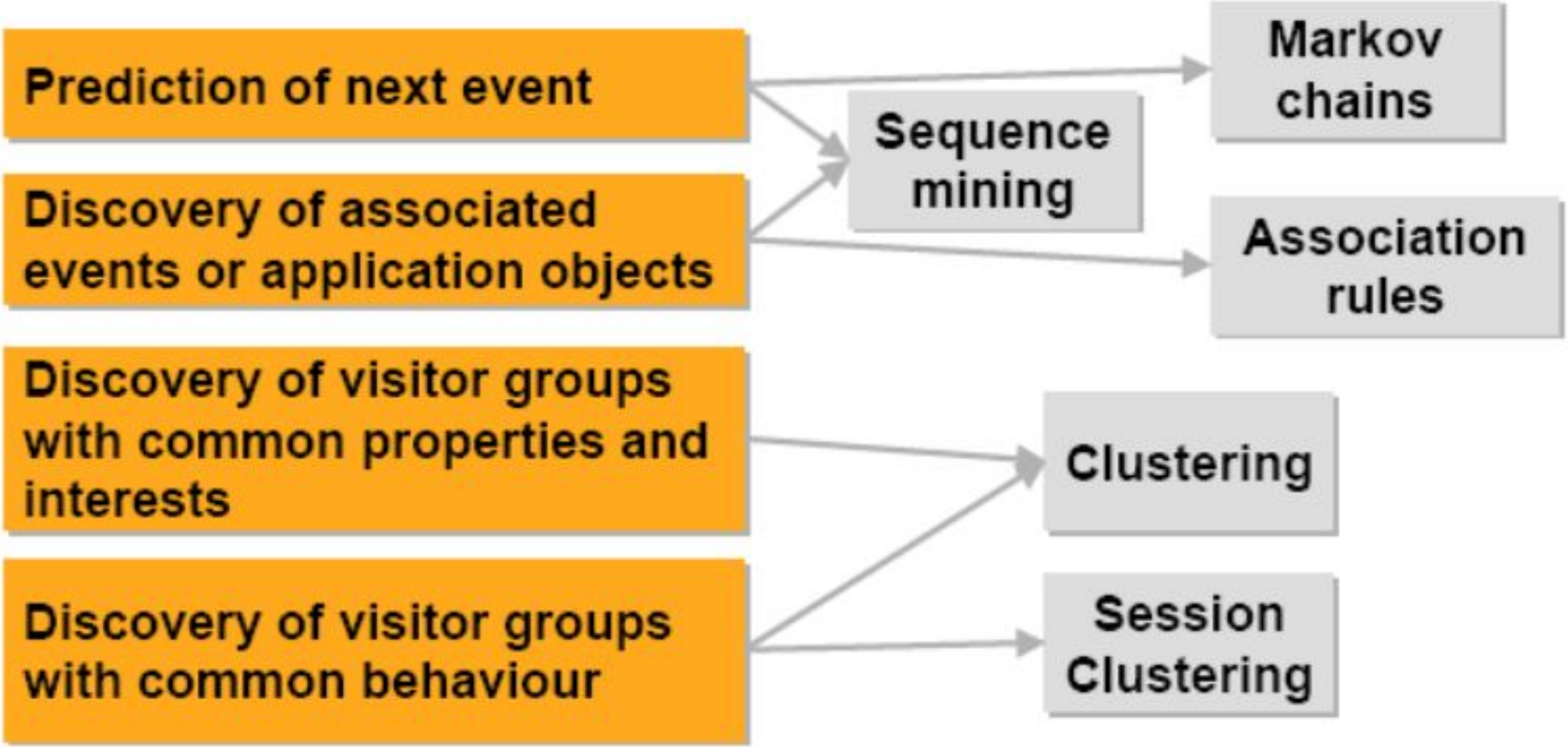
Prepared by:
U17CO044 Suryansh Singh Rathore

# What is Clickstream data?

Clickstream is a sequence of events that represent a visitor’s actions on a website. It may include collecting clicks, purchases, views, impressions and any other event relevant to the business. Clickstreams are among the most popular data sources because web servers automatically record each action in their web log entries.

# Why Clickstream is so important?

- To know about visitors/customers
- Personalize their product experience
- Data is important for a business to stay competitive
- Analyzing employee productivity
- Software testing
- General research
- Selective advertising



Prediction of next event

Discovery of associated events or application objects

Discovery of visitor groups with common properties and interests

Discovery of visitor groups with common behaviour

Markov chains

Sequence mining

Association rules

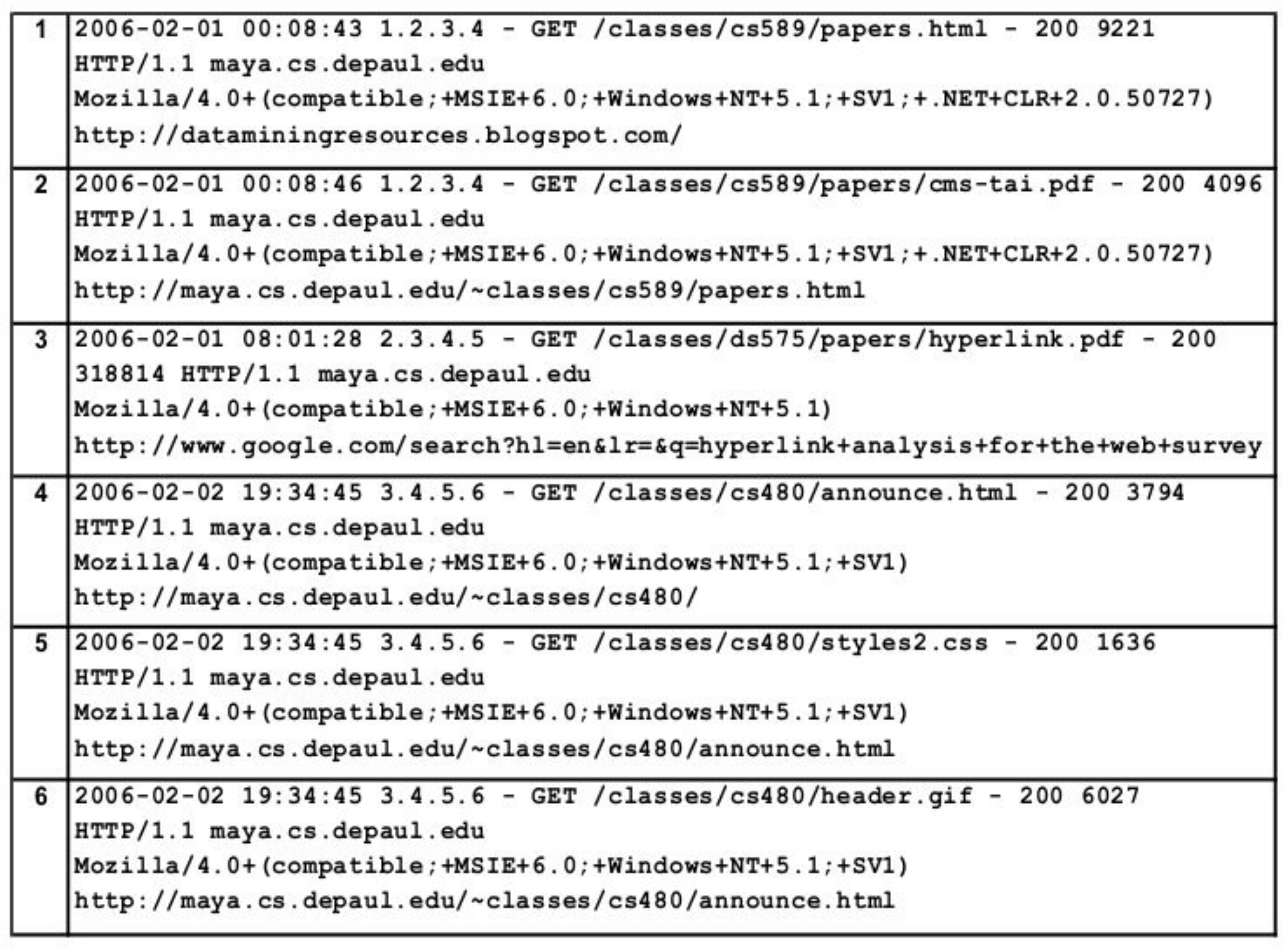
Clustering

Session Clustering

# Sample entry in a log file of web server:

This log entry contains:

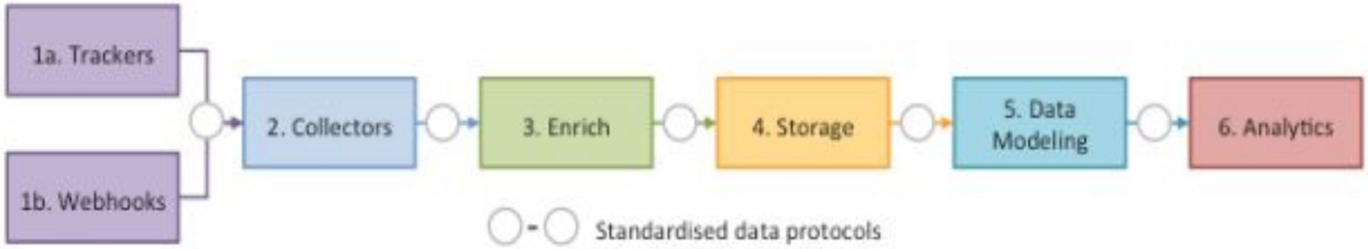
- IP address of visitor, possibly cookie ID
- Date and time of page request
- Returned HTTP status code, number of bytes transferred from website
- A precise HTTP request
- HTTP methods
- The most recent referring URL
- The requesting system information



| | |
|---|---|
| 1 | 2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://dataminingresources.blogspot.com/ |
| 2 | 2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://maya.cs.depaul.edu/~classes/cs589/papers.html |
| 3 | 2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200<br>318814 HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)<br>http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey |
| 4 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/ |
| 5 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/announce.html |
| 6 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/announce.html |

# Clickstream data mining process:

Mostly Clickstream data is collected using Javascript tracker scripts which is loaded with page on every request. These data are the sessions of the site visitors.



Event transformation real time data pipeline –

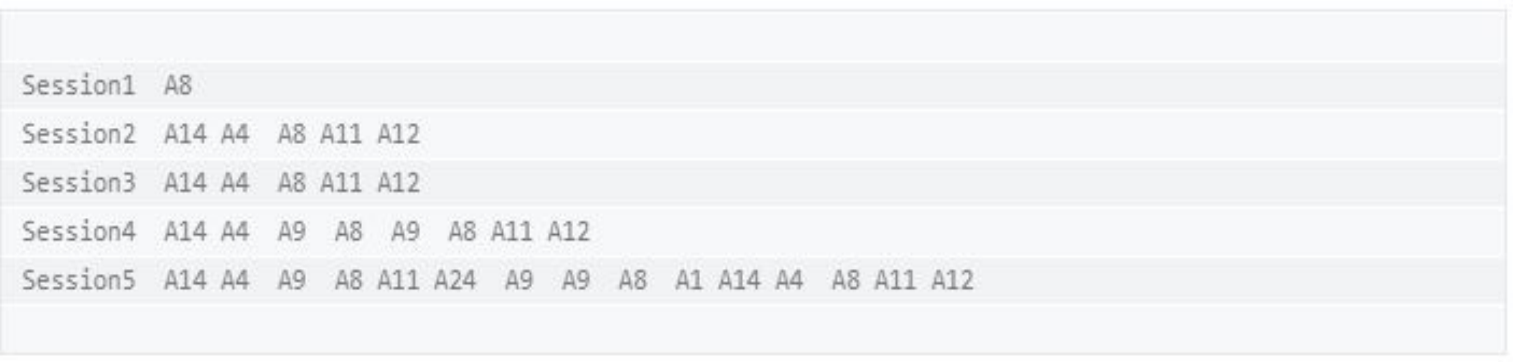
Image credit: https://github.com/snowplow/snowplow (https://github.com/snowplow/snowplow)

# Clickstream data mining process (Cont’d):

- Extraction: Web activities (add ratings, add reviews etc), Forms, Cookies, Javascript tracker scripts
- Enrich: Data cleaning, Sessionization, Data integration, Data transformation
- Storage: User transaction database (Customer profiling)
- Data modeling: Pageview clustering, Correlation analysis, Association rule mining, Sequential pattern mining etc
- Analytics: Pattern filtering, Aggregation, Categorization

# Sessionization:

- Identifying sessions for user identification
- Split the log and identify sessions
- The time frame from the moment a user enters the site until the moment he/she leaves it is a session
- Why sessionization? - proxy servers, dynamic IP addresses, missing references due to caching



| | |
|---|---|
| Session1 | A8 |
| Session2 | A14 A4 A8 A11 A12 |
| Session3 | A14 A4 A8 A11 A12 |
| Session4 | A14 A4 A9 A8 A9 A8 A11 A12 |
| Session5 | A14 A4 A9 A8 A11 A24 A9 A9 A8 A1 A14 A4 A8 A11 A12 |

| Time | IP | URL | Ref | Agent |
|---|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - | IE5;Win2k |
| 0:09 | 1.2.3.4 | B | A | IE5;Win2k |
| 0:10 | 2.3.4.5 | C | - | IE6;WinXP;SP1 |
| 0:12 | 2.3.4.5 | B | C | IE6;WinXP;SP1 |
| 0:15 | 2.3.4.5 | E | C | IE6;WinXP;SP1 |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:22 | 2.3.4.5 | D | B | IE6;WinXP;SP1 |
| 0:22 | 1.2.3.4 | A | - | IE6;WinXP;SP2 |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 0:25 | 1.2.3.4 | C | A | IE6;WinXP;SP2 |
| 0:33 | 1.2.3.4 | B | C | IE6;WinXP;SP2 |
| 0:58 | 1.2.3.4 | D | B | IE6;WinXP;SP2 |
| 1:10 | 1.2.3.4 | E | D | IE6;WinXP;SP2 |
| 1:15 | 1.2.3.4 | A | - | IE5;Win2k |
| 1:16 | 1.2.3.4 | C | A | IE5;Win2k |
| 1:17 | 1.2.3.4 | F | C | IE6;WinXP;SP2 |
| 1:26 | 1.2.3.4 | F | C | IE5;Win2k |
| 1:30 | 1.2.3.4 | B | A | IE5;Win2k |
| 1:36 | 1.2.3.4 | D | B | IE5;Win2k |

**User 1**

| | | | |
|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - |
| 0:09 | 1.2.3.4 | B | A |
| 0:19 | 1.2.3.4 | C | A |
| 0:25 | 1.2.3.4 | E | C |
| 1:15 | 1.2.3.4 | A | - |
| 1:26 | 1.2.3.4 | F | C |
| 1:30 | 1.2.3.4 | B | A |
| 1:36 | 1.2.3.4 | D | B |

**User 2**

| | | | |
|---|---|---|---|
| 0:10 | 2.3.4.5 | C | - |
| 0:12 | 2.3.4.5 | B | C |
| 0:15 | 2.3.4.5 | E | C |
| 0:22 | 2.3.4.5 | D | B |

**User 3**

| | | | |
|---|---|---|---|
| 0:22 | 1.2.3.4 | A | - |
| 0:25 | 1.2.3.4 | C | A |
| 0:33 | 1.2.3.4 | B | C |
| 0:58 | 1.2.3.4 | D | B |
| 1:10 | 1.2.3.4 | E | D |
| 1:17 | 1.2.3.4 | F | C |

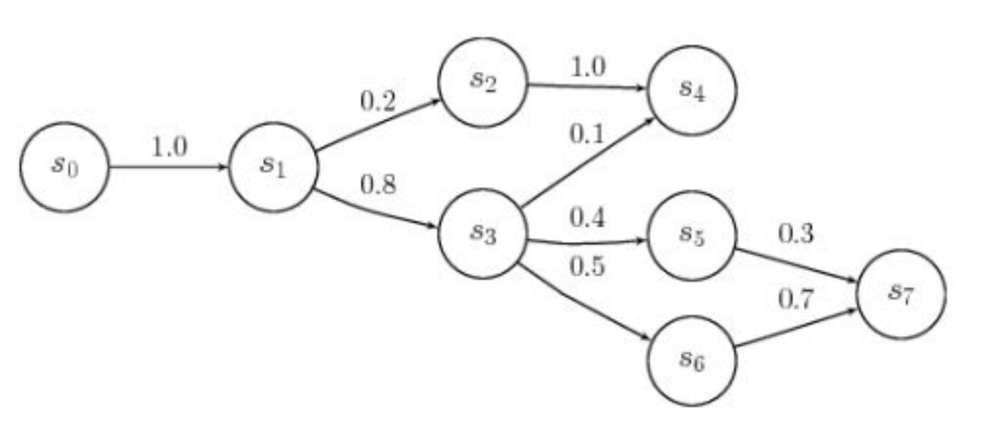
# Standard approaches for Customer Profiling:

- **Rule-based filtering:** content provided/suggested based on predefined rules (e.g. “if user has clicked on A and has location B, has age <C then recommend D”)
- **Collaborative filtering:** give recommendations based on response/ratings of other “similar” users
- **Content-based filtering:**  track which pages the user visits and recommend other pages with similar content
- **Hybrid methods:** usually combination of content-based and collaborative

# Markov Chain algorithm for mining Clickstream data:

Markov chains work best with sequential data. It is a stochastic process. It is a random process in which future is independent of past, given the present.

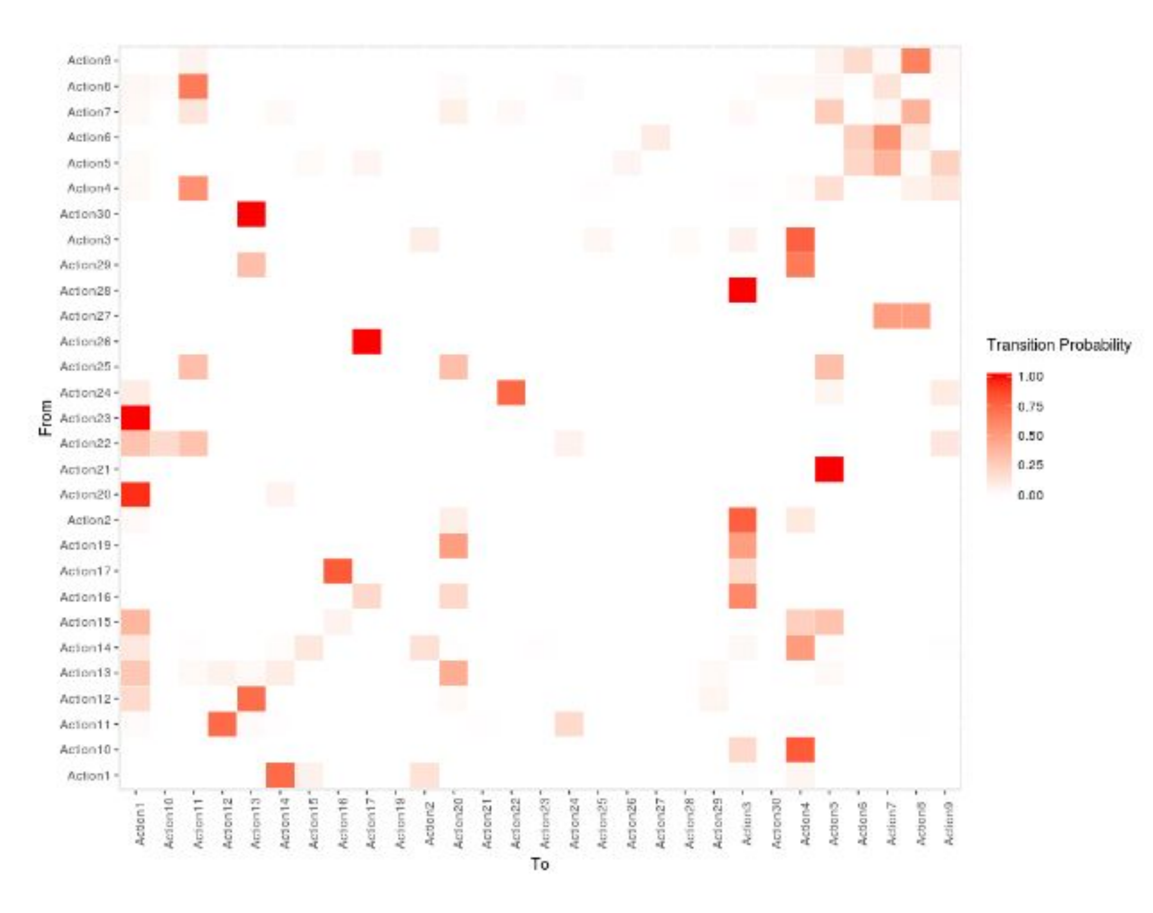
Sample Markov chain:



# Markov Chain algorithm (Cont'd):

- The order of the Markov chain is said to be the number states on which the current state depends.
- Hence a zero-order chain implies that probability of being in a state is completely independent on any and all previous states.
- Higher-order Markov chain gives more realistic models but increases complexity.
- Sequence of actions taken by a user on the website are stored within a session.
- The website log file contains many such sessions info
- Based on this data we build transition probability matrix, which represents the probability of an action to be occured based on current state using our Markov Chain

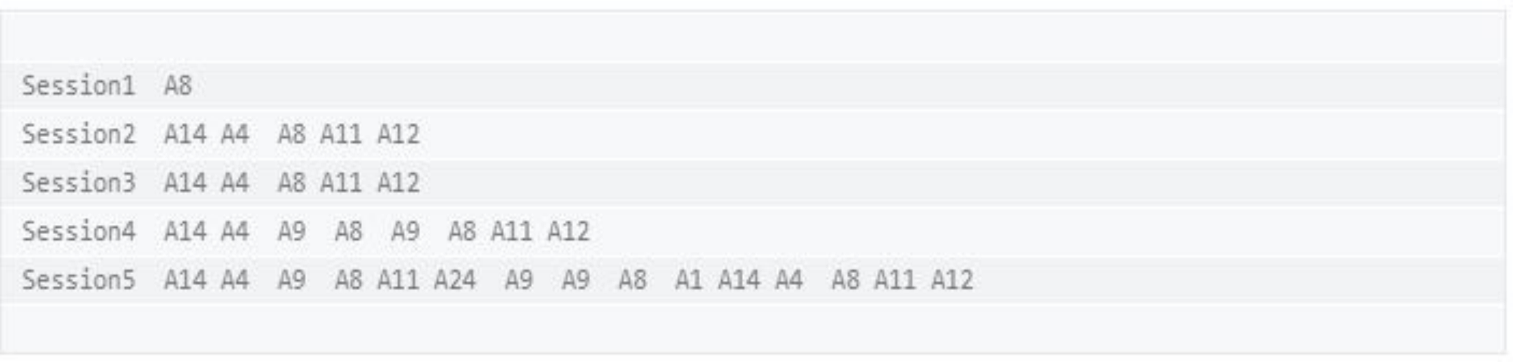
# Transition probability Heat Map:



# cSPADE algorithm for mining Clickstream data:

The cSPADE algorithm can be used to explore, understand and even predict a given customer’s navigation patterns, unlike Markov Chains algorithm.

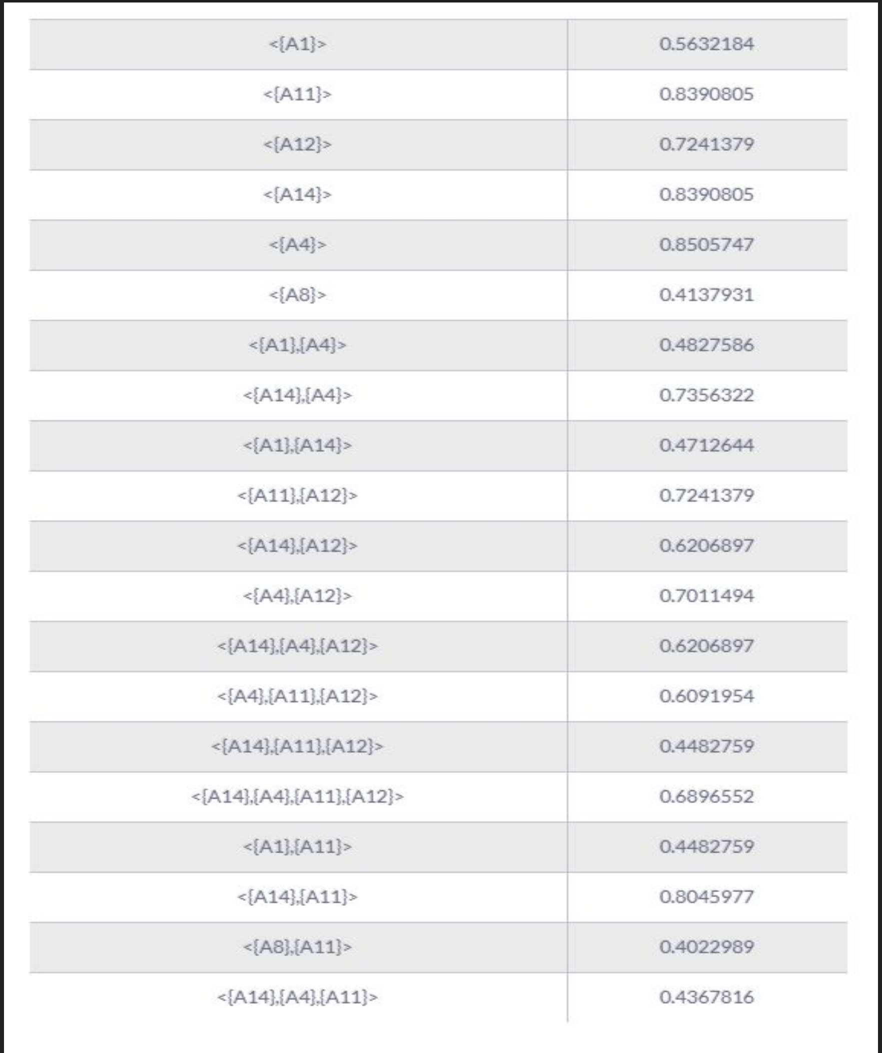
Consider this dataset:



```
Session1  A8
Session2  A14 A4  A8 A11 A12
Session3  A14 A4  A8 A11 A12
Session4  A14 A4  A9  A8  A9  A8 A11 A12
Session5  A14 A4  A9  A8 A11 A24  A9  A9  A8  A1 A14 A4  A8 A11 A12
```

# cSPADE algorithm (Cont’d):

- In the first step, algorithm calculates the probability of occurring first actions.
- In the second step, it computes frequency of occurring two actions one after the other.
- Similarly for three, four and so on….
- The sequence of actions and their respective probabilities for a sample dataset looks like:



| | |
|---|---|
| <{A1}> | 0.5632184 |
| <{A11}> | 0.8390805 |
| <{A12}> | 0.7241379 |
| <{A14}> | 0.8390805 |
| <{A4}> | 0.8505747 |
| <{A8}> | 0.4137931 |
| <{A1},{A4}> | 0.4827586 |
| <{A14},{A4}> | 0.7356322 |
| <{A1},{A14}> | 0.4712644 |
| <{A11},{A12}> | 0.7241379 |
| <{A14},{A12}> | 0.6206897 |
| <{A4},{A12}> | 0.7011494 |
| <{A14},{A4},{A12}> | 0.6206897 |
| <{A4},{A11},{A12}> | 0.6091954 |
| <{A14},{A11},{A12}> | 0.4482759 |
| <{A14},{A4},{A11},{A12}> | 0.6896552 |
| <{A1},{A11}> | 0.4482759 |
| <{A14},{A11}> | 0.8045977 |
| <{A8},{A11}> | 0.4022989 |
| <{A14},{A4},{A11}> | 0.4367816 |

From the above table, we see that for a given sequence pattern X, it is possible to predict the next click by searching for the pattern sequence with the highest support starting with X.
For example, after performing the action A14, the most probable next action is A11, according to the pattern sequence 8 – with a probability of 0.8045.

Some other clickstream data modeling techniques:
Traffic analysis, Sales funnel analysis, Browse/Cart abandonment and recovery, A/B testing, Identity stitching, RFM analysis.

# Popular Clickstream data vendors:

Mixpanel, Google 360, Adobe Marketing Cloud, Google Analytics, Snowplow, Kissmetrics, Amplitude, Heap, Matomo etc.

Mostly Clickstream data vendors provide a Javascript tracker script that is attached directly to a target website or included in tag manager bundles.

# References:

- https://www.slideshare.net/ERSHUBHAMTIWARI/clickstream-analysis-57968355
- https://rudderstack.com/blog/data-mining-for-clickstream-analytics/
- https://www.zora.uzh.ch/id/eprint/56367/1/56367.pdf
- https://stacktome.com/blog/a-guide-to-data-warehousing-clickstream-data
- https://www.dais.unive.it/~dm/New_Slides/10_WUM.pdf

# Thank You