Analyzing and Visualizating Event Sequence Data

Sean J. Taylor
Chief Scientist

motif


January 31st, 2013
Pivotal Labs
841 Broadway
Jared Lander
Me
Wes McKinney
John Myles White
Photo Credit:
Drew Conway

Sean

6649d

Graduate_from_Penn

61d

Start_job_at_the_Fed

761d

Start_job_at_Matrix_Group

731d

Began_PhD_at_NYU

1403d

Internship_at_Facebook

241d

Talk_at_NYR_Meetup

117d

event name	Talk_at_NYR_Meetup
person actor	Sean
timestamp ts	2013-01-31 00:00:00

Move_to_San_Francisco

157d

Start_job_at_Facebook

164d

Graduate_from_NYU

1857d

Leave_job_at_Facebook

5d

Start_job_at_Lyft

1016d

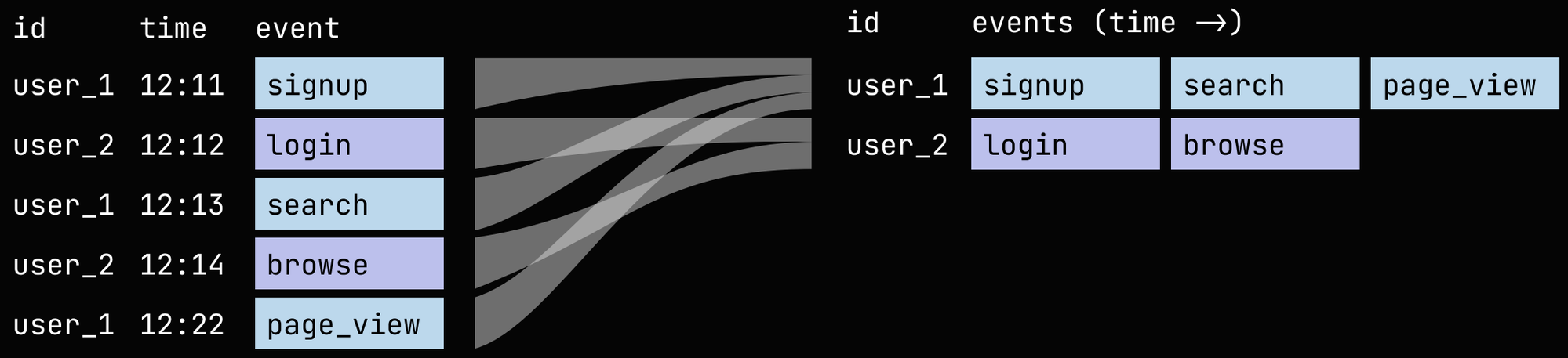
Leave_job_at_Lyft

136d

Co_found_Motif_Analytics

event name	Co_found_Motif_Analytics
person actor	Sean
timestamp ts	2022-07-15 00:00:00

Event Sequence Data



Very common form of analytics data with two properties per event:

1. **Time dimension**: events have precise timestamps.

2. **Cross-sectional dimension**: usually user ID.

If you are:

- **An Economist**: this is a more general case of panel data.

- **A Data engineer**: this is a transactional table, the most fundamental type of fact table.

Data-Specific Approaches

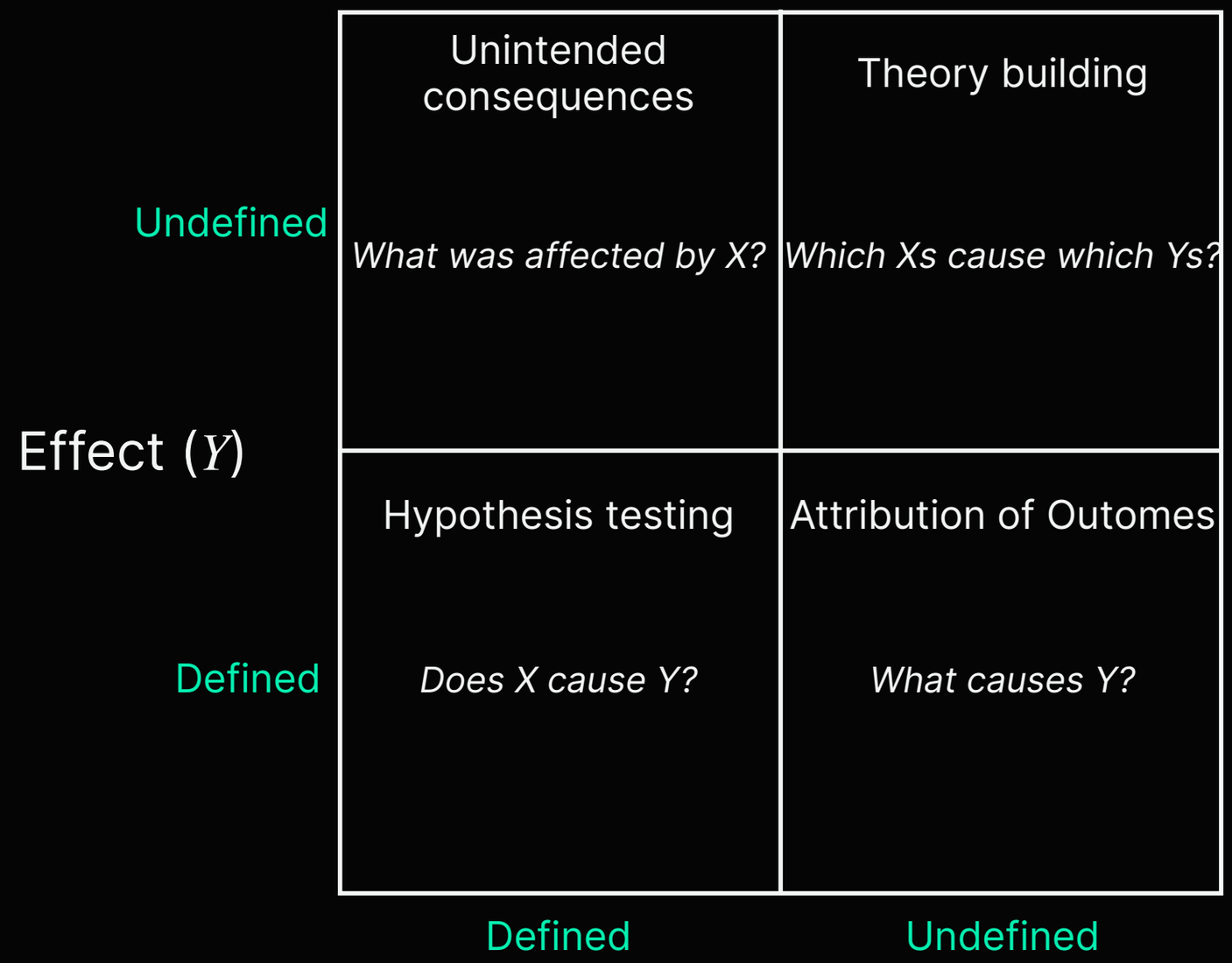
Data with known structure

- Time series
- Graphs
- Text
- Images
- Spatial data
- Panel data
- Event Sequences

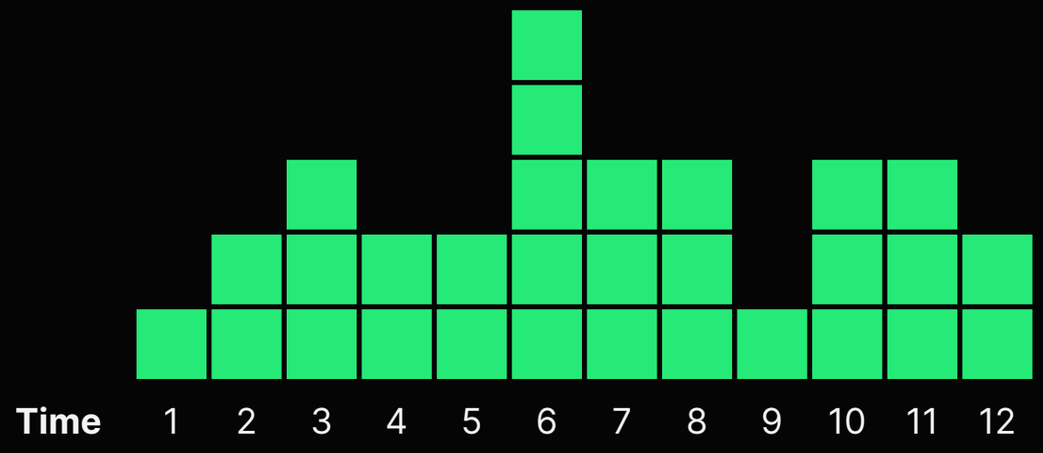
Tools leverage structure

- Encoding / Decoding
- Storage
- Cleaning and checking
- Querying
- Analysis
- Modeling
- Visualization

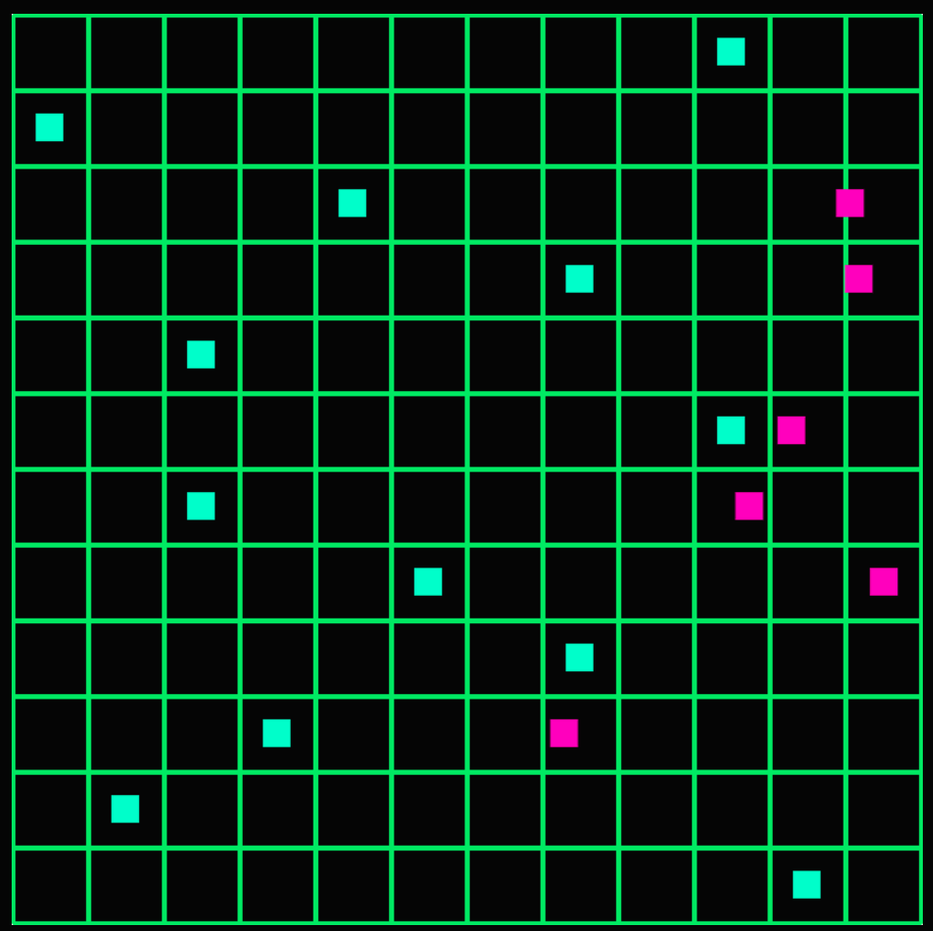
Analytics Tasks



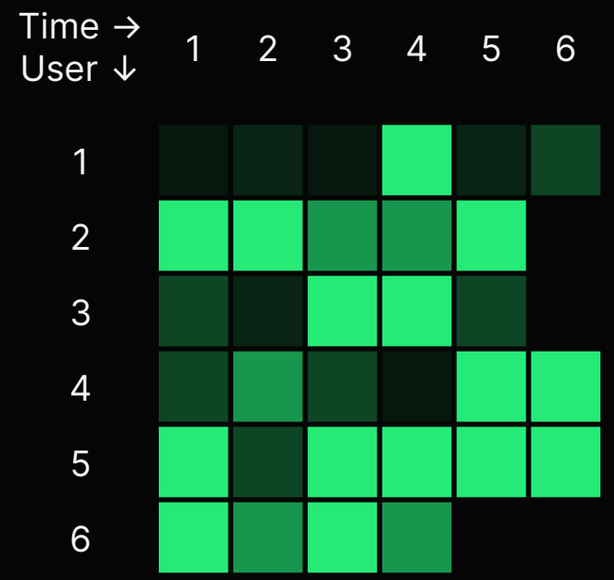
- **Defined**: in the data and with a agreed upon measurement strategy.
- **Undefined**: possibly in the data but without a specified definition or measurement strategy.



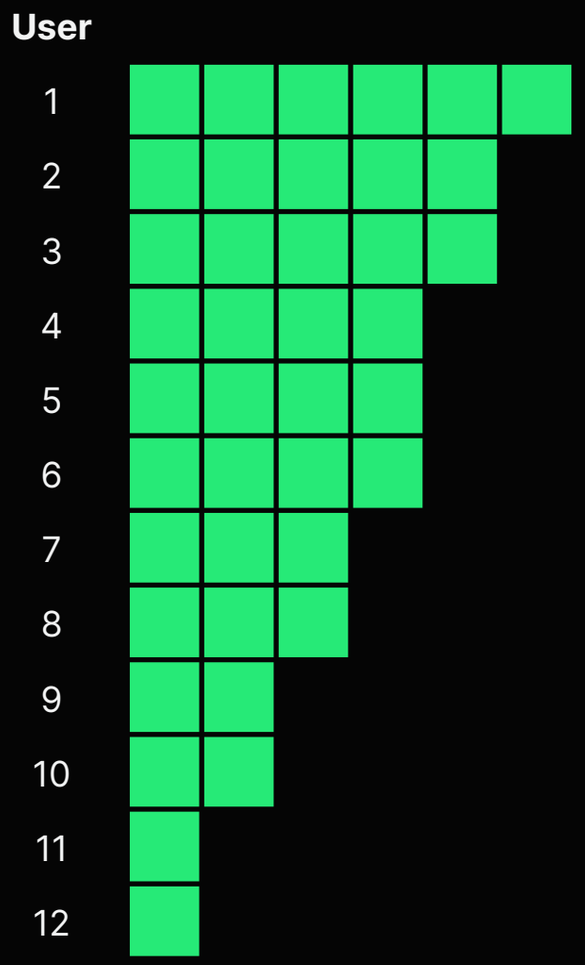
Time Aggregation



User x Time

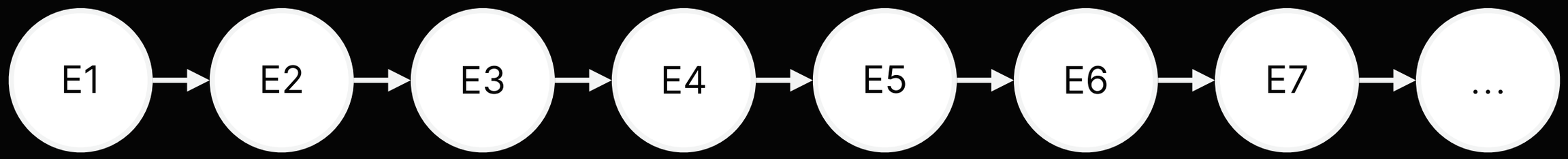


User Aggregation



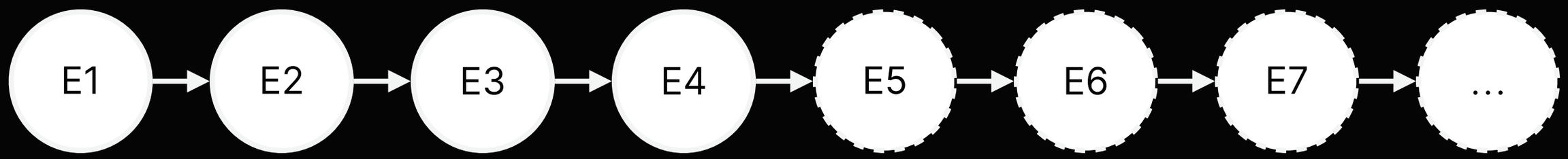
Event Sequence Models

Observed Event Sequence



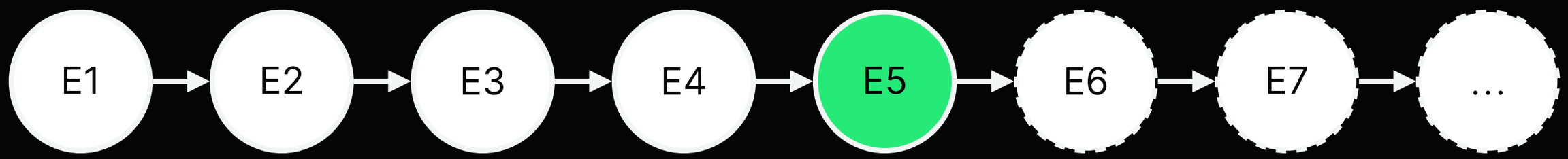
Conditional Sequence Distribution

$$\Pr(E_{5:} \mid E_{1:4})$$



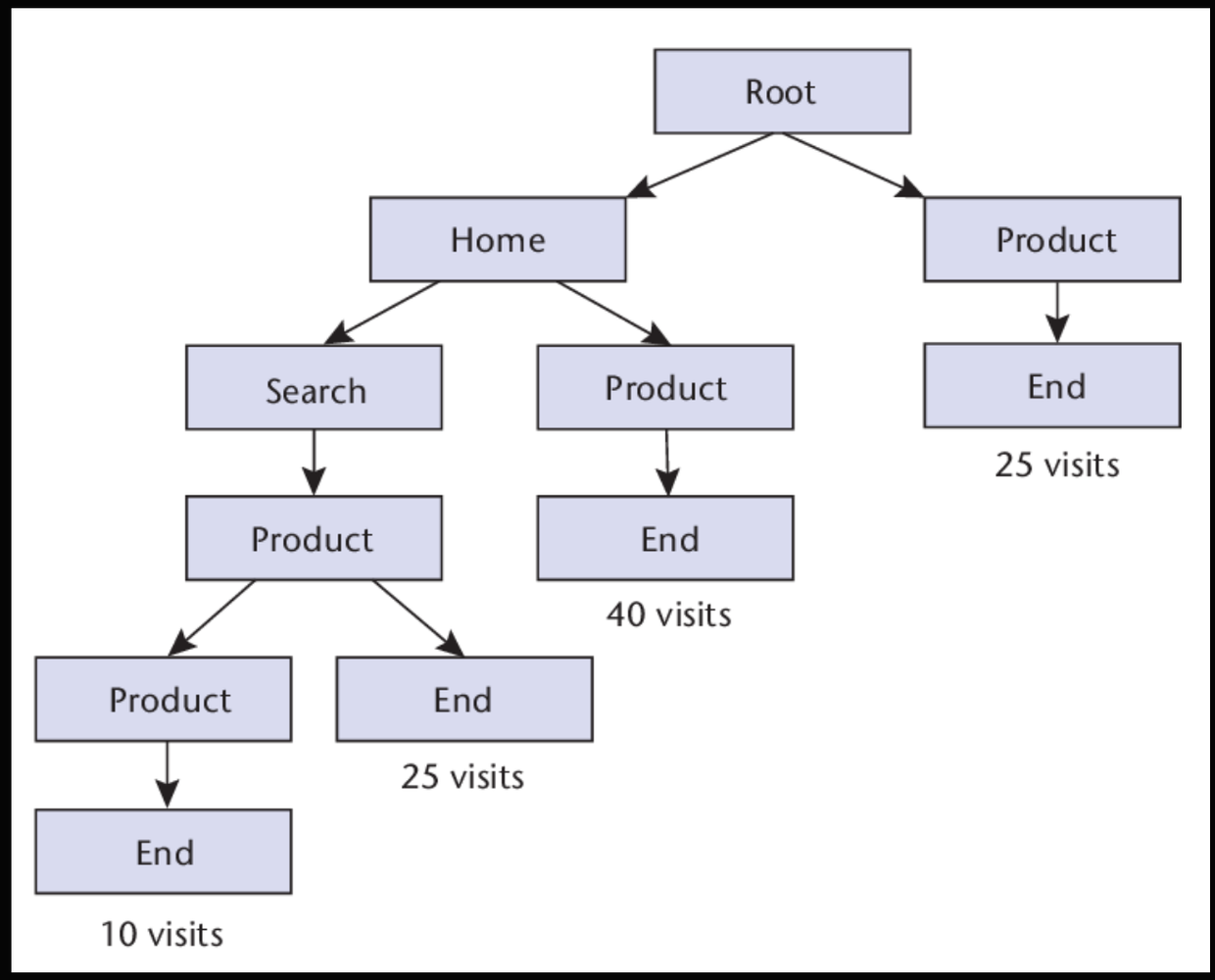
Interventional Distribution

$$\Pr(E_{6:} \mid \mathbf{do}(E_5), E_{1:4})$$



Prefix Tree (Trie)

- Each node represents a prefix of a sequence (special node for start).
- Increment counter for each sequence with a matching prefix.
- **Number of parameters** grows exponentially in sequence length and variety.
- Control complexity by **shortening sequences** in a deterministic way: omit events, truncate, split, etc.



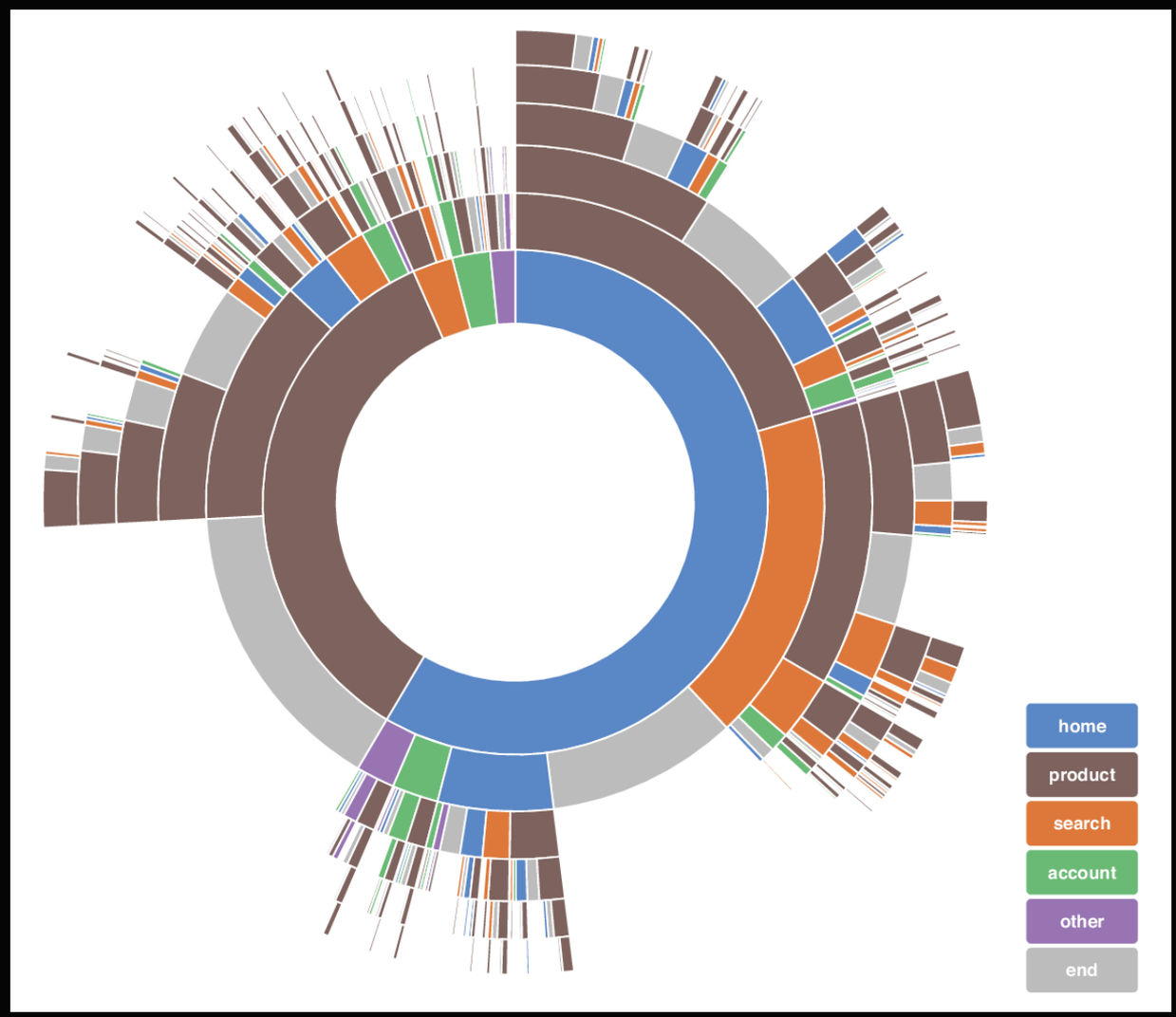
Tree Visualization

Radius: position in sequence

Angle: probability of event

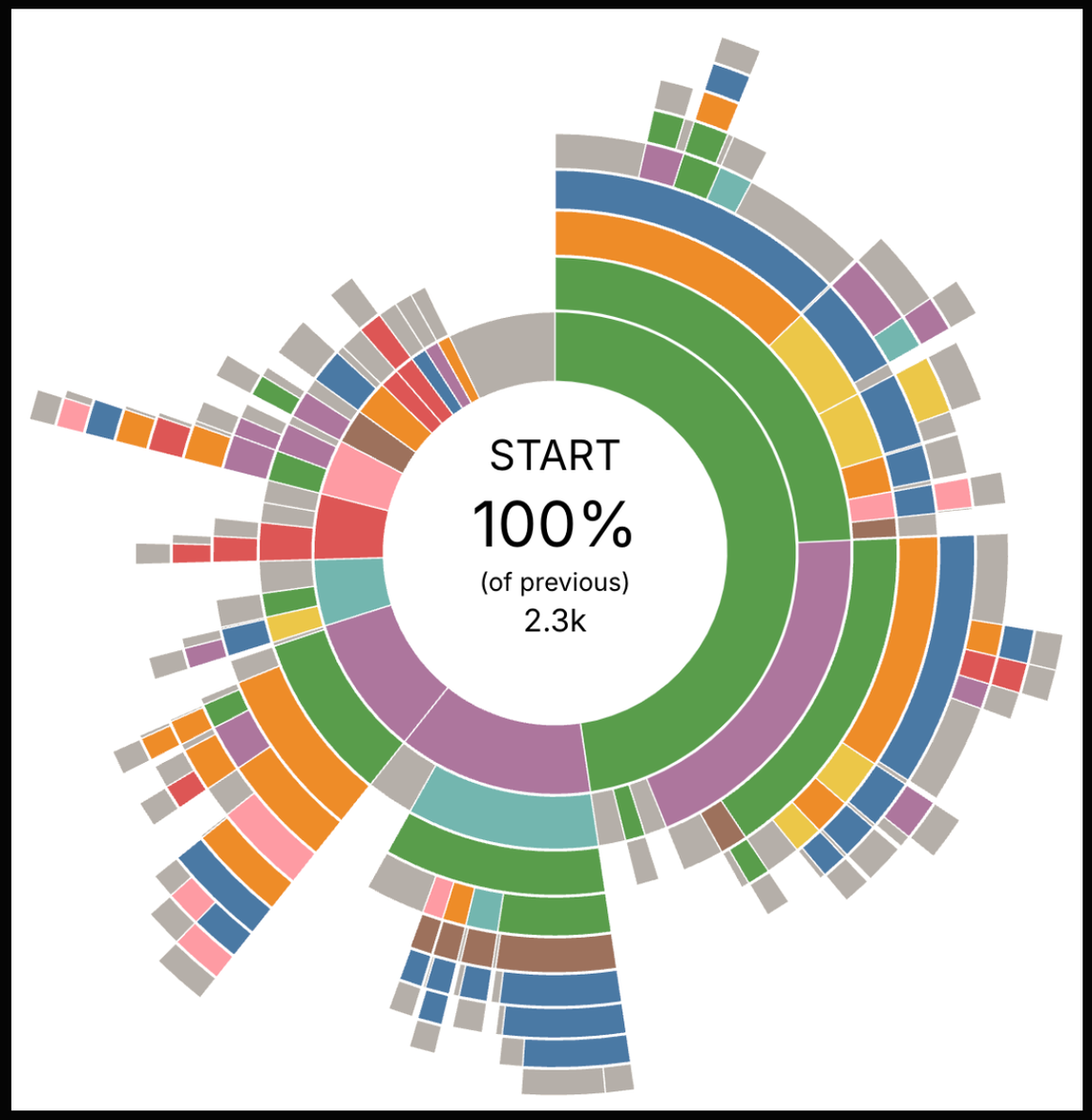
Color: event type

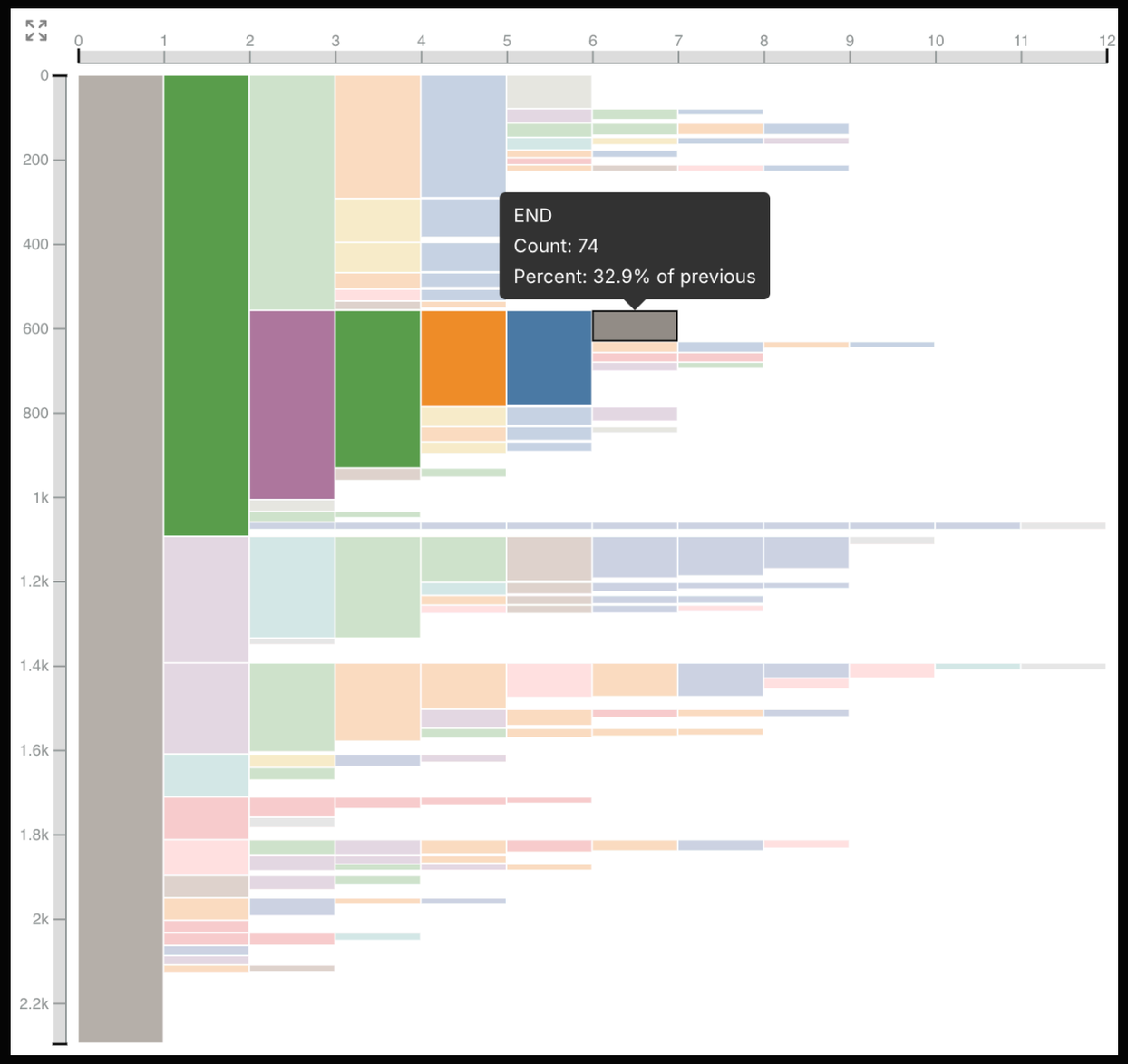
- **Condition** on prefixes by clicking on specific segments.
- **Scalability**: sequences must be short, often less than length 10.
- *Think of it as a special histogram for sequences of categorical variables.*



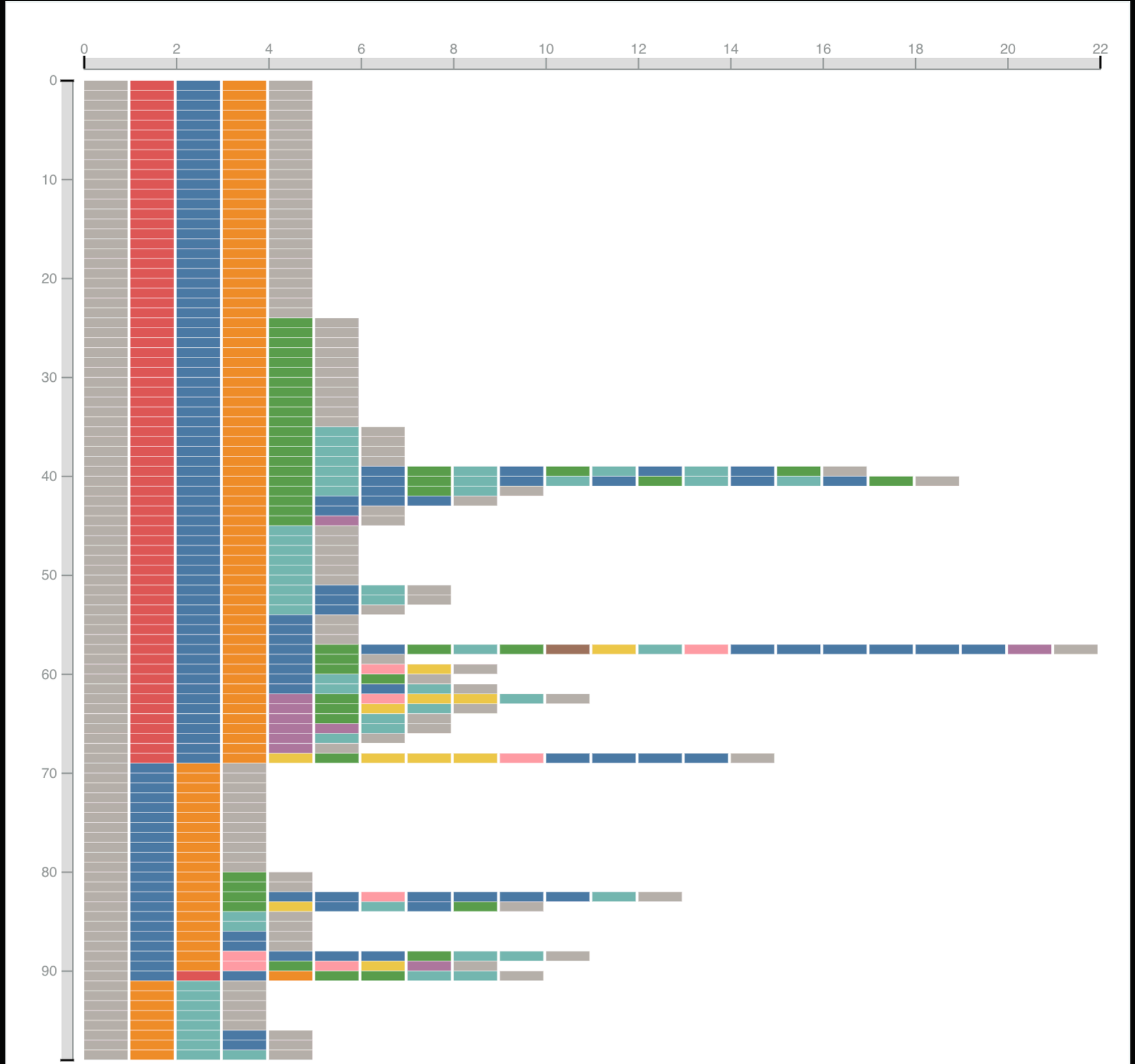
Sunburst Visualization (Rodden, 2014)

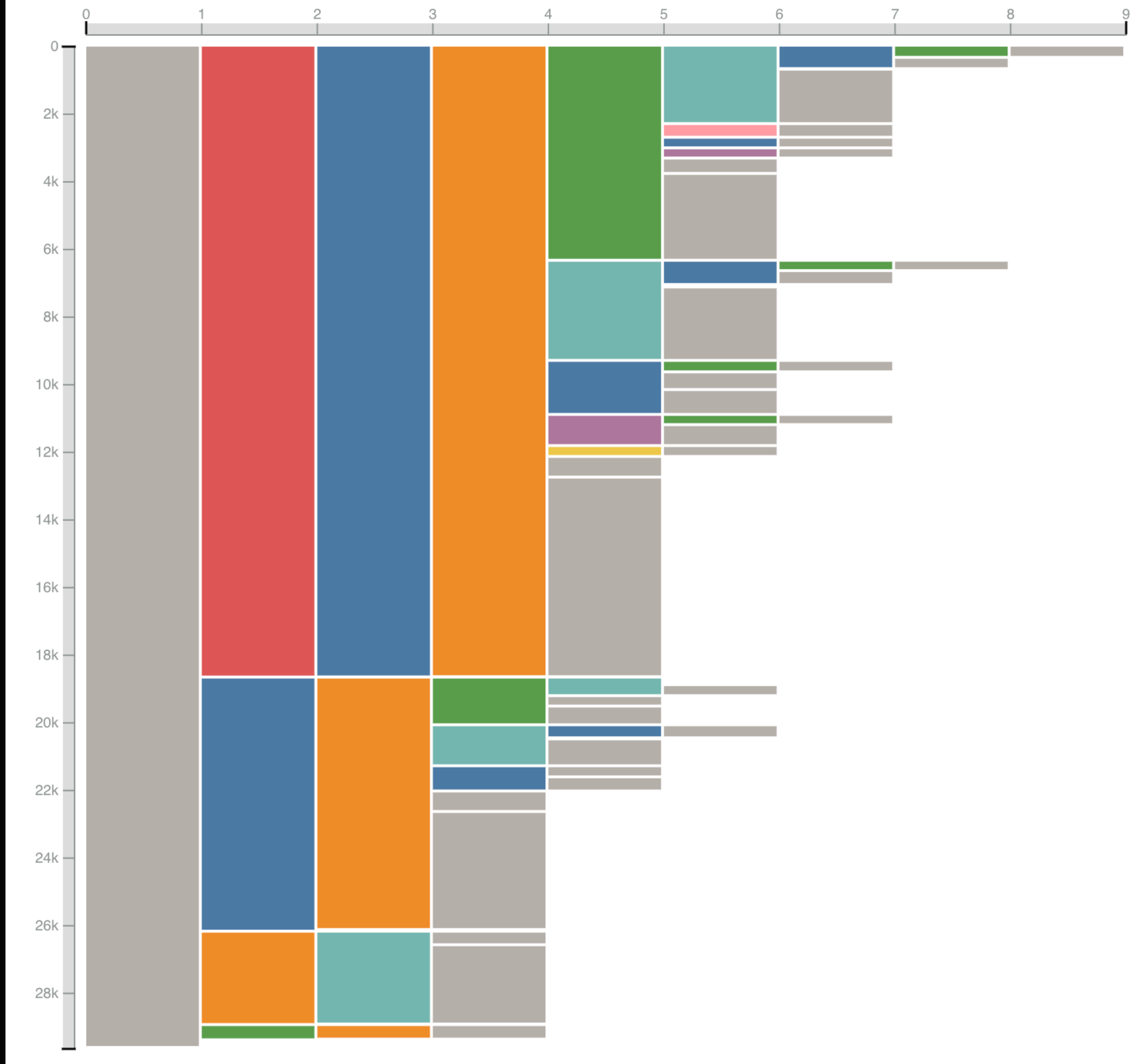
Radial → Linear





Actor event_para...	Sequence									
1707873321	page_view	session_start								
1700175998	page_view	session_start								
1707991182	first_visit	page_view	session_start							
1711481663	first_visit	page_view	session_start							
1707820669	first_visit	page_view	session_start							
1714445272	first_visit	page_view	session_start	user_engagem...	page_view	user_engagem...				
1714152877	first_visit	page_view	session_start	user_engagem...						
1708454364	first_visit	page_view	session_start	scroll	user_engagem...	page_view	scroll	user_engagem...	page_view	scroll
1715165797	first_visit	page_view	session_start	scroll	user_engagem...					
1701374302	first_visit	page_view	session_start	click	scroll	click	click	click	blogClick	page_view
1710313077	page_view	session_start								
1689636172	first_visit	page_view	session_start	playVideo	scroll	blogClick	click	click	user_engagem...	
1685116480	first_visit	page_view	session_start	scroll	playVideo					
1701342583	page_view	session_start								
1703883561	first_visit	page_view	session_start	scroll	page_view	page_view	page_view			
1707956420	first_visit	page_view	session_start	page_view						

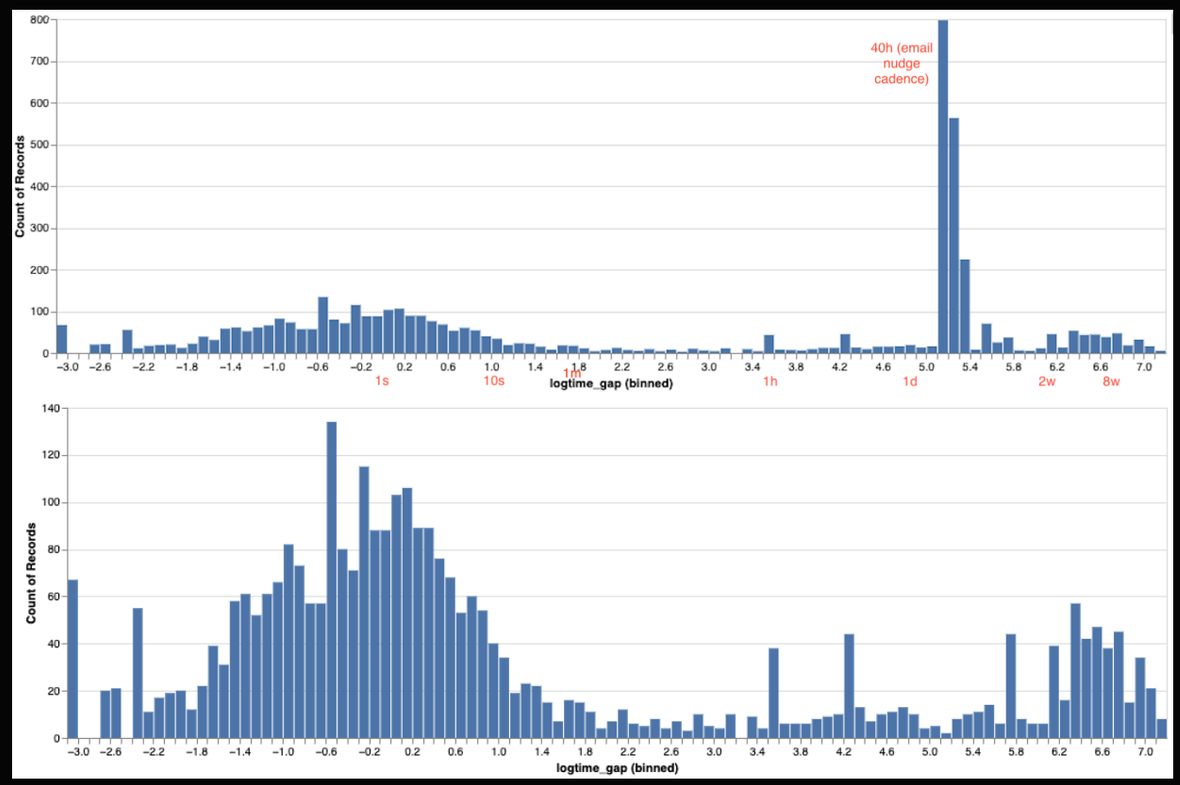

0
2
4
6
8
10
12
14
16
18
20
22
0
10
20
30
40
50
60
70
80
90


0
1
2
3
4
5
6
7
8
9
0
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
22k
24k
26k
28k

Sequence Simplification

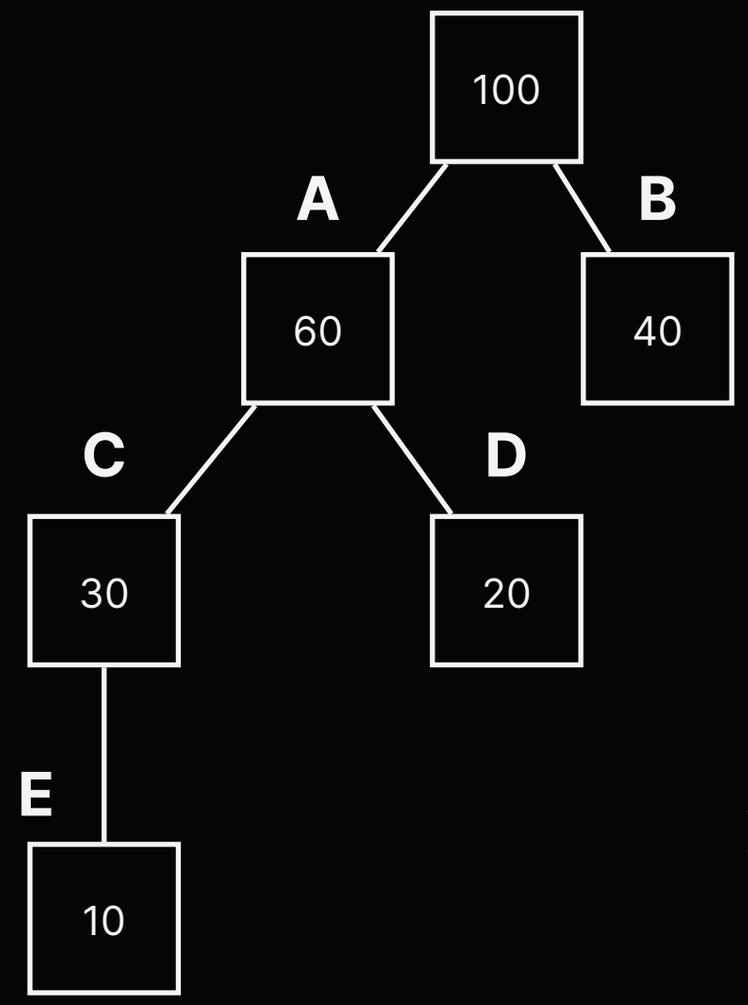
- **Removing events**: reduce noise by omitting events that are not meaningful.
- **Merging events**: replace contiguous sets of events with summaries.
- **Truncation**: setting some maximum time bounds the length of the sequences.
- **Splitting sequences**: focus analysis on activity within a contiguous time period. (e.g. sessionization)
- Redefining the start of the sequence changes the definition of $\Pr(S)$!

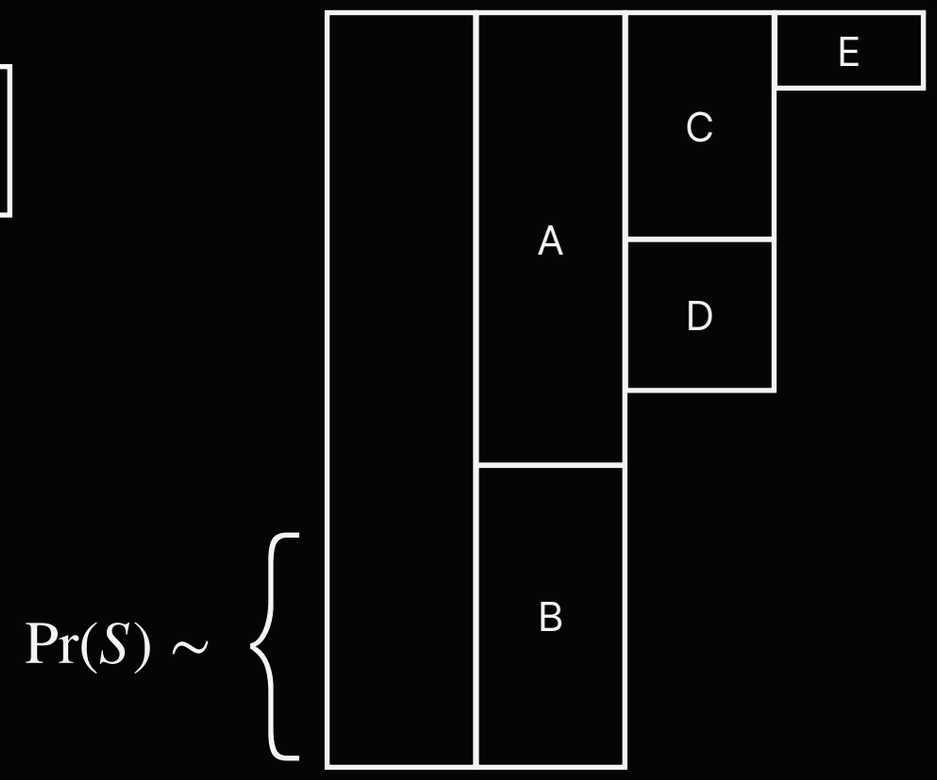
Distribution of time between events



$\Pr(S)$ when S is a Sequence

Sequence	Count
A	10
B	40
A→C	20
A→C→E	10
A→D	20





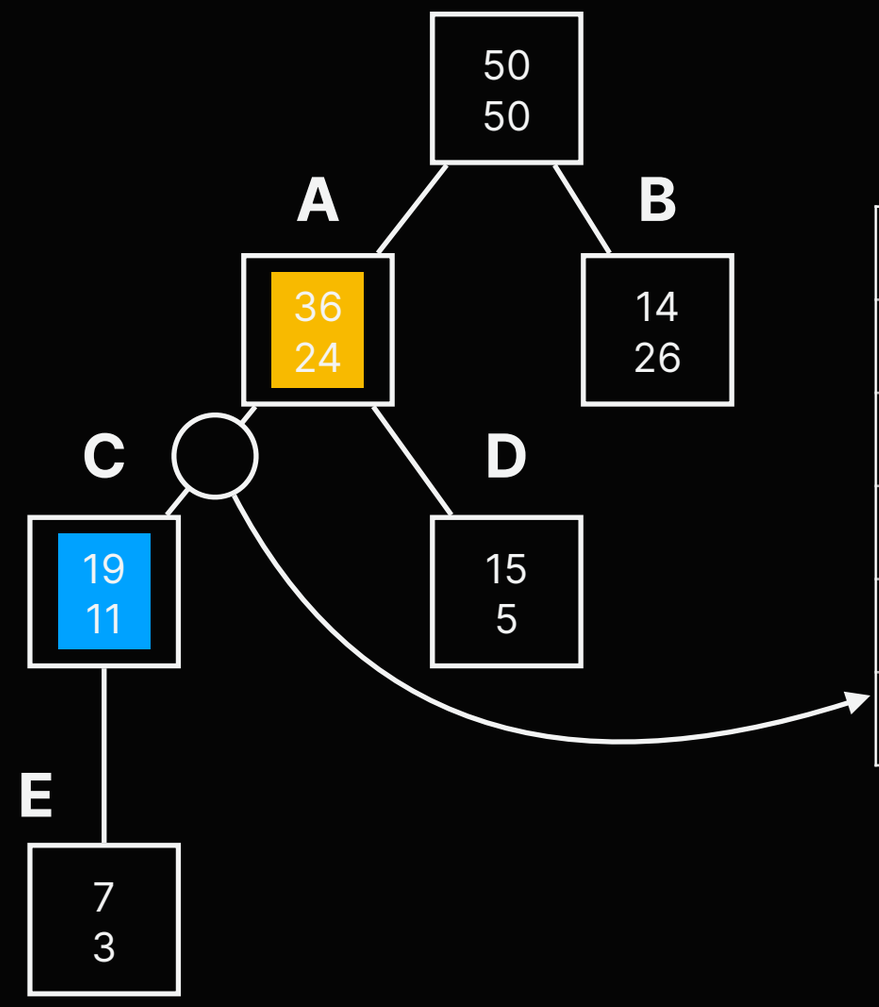
Raw data

Tree representation

Visualization

$\Pr(S \mid X)$ when $X \in \{0,1\}$

Sequence	X=0	X=1
A	2	8
B	14	26
A→C	12	8
A→C→E	7	3
A→D	15	5



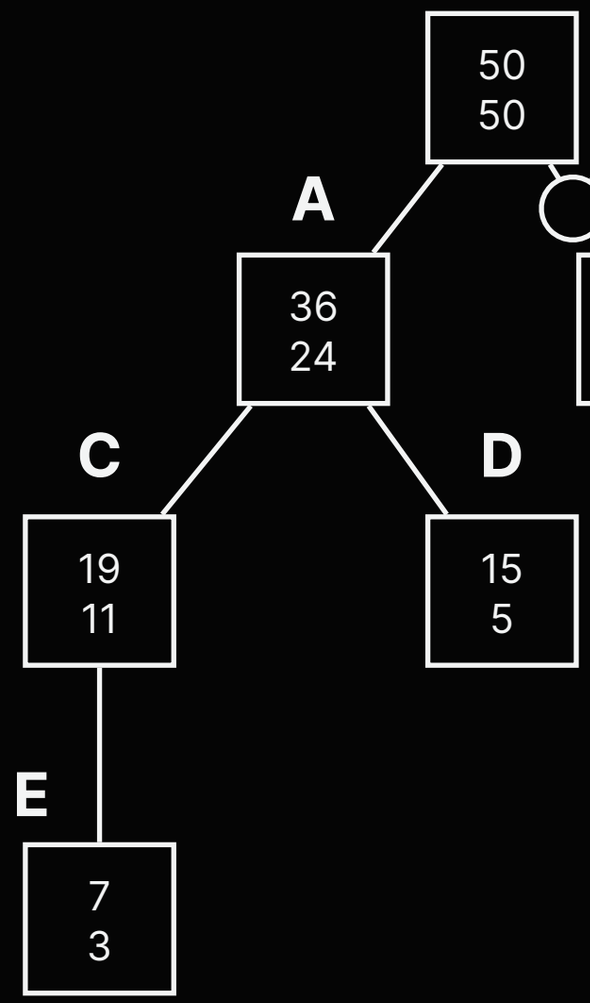
	X=0	X=1	Δ
A	36	24	
A→C	19	11	
(A→C) / (A)	53%	46%	-7%
Standard Error	8%	11%	13%
P-Value			0.87

Raw data
e.g. from A/B test

Tree representation
two counts per node

Edge Statistic
difference in transition rates

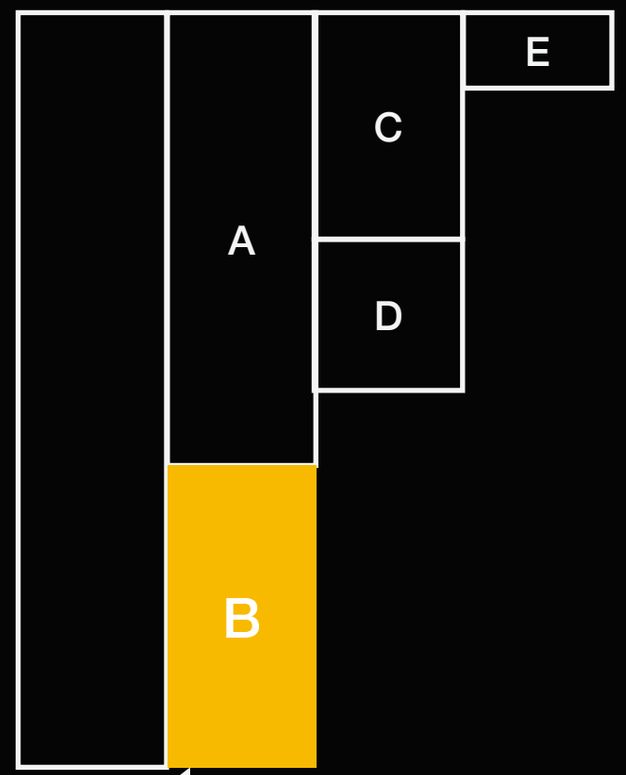
Comparison Visualization

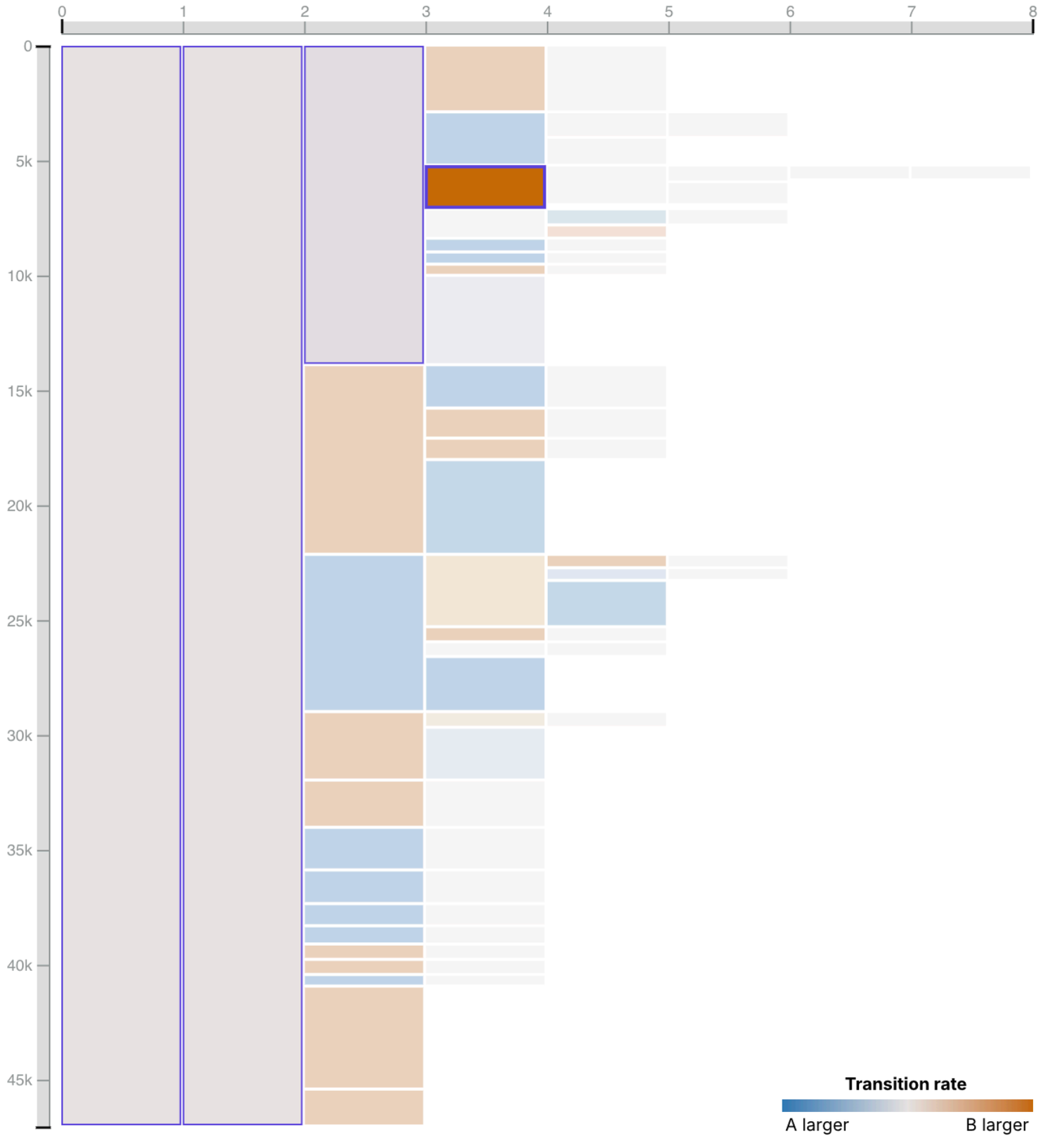


We do a statistical test for each edge in the tree

	X=0	X=1	
Start	50	50	
Start→B	14	26	
(Start→B) / (Start)	28%	52%	14%
Standard Error	8%	7%	10%
P-Value			0.02

Significant Differences
(more saturated colors)





>> Commute ×

View example sequence

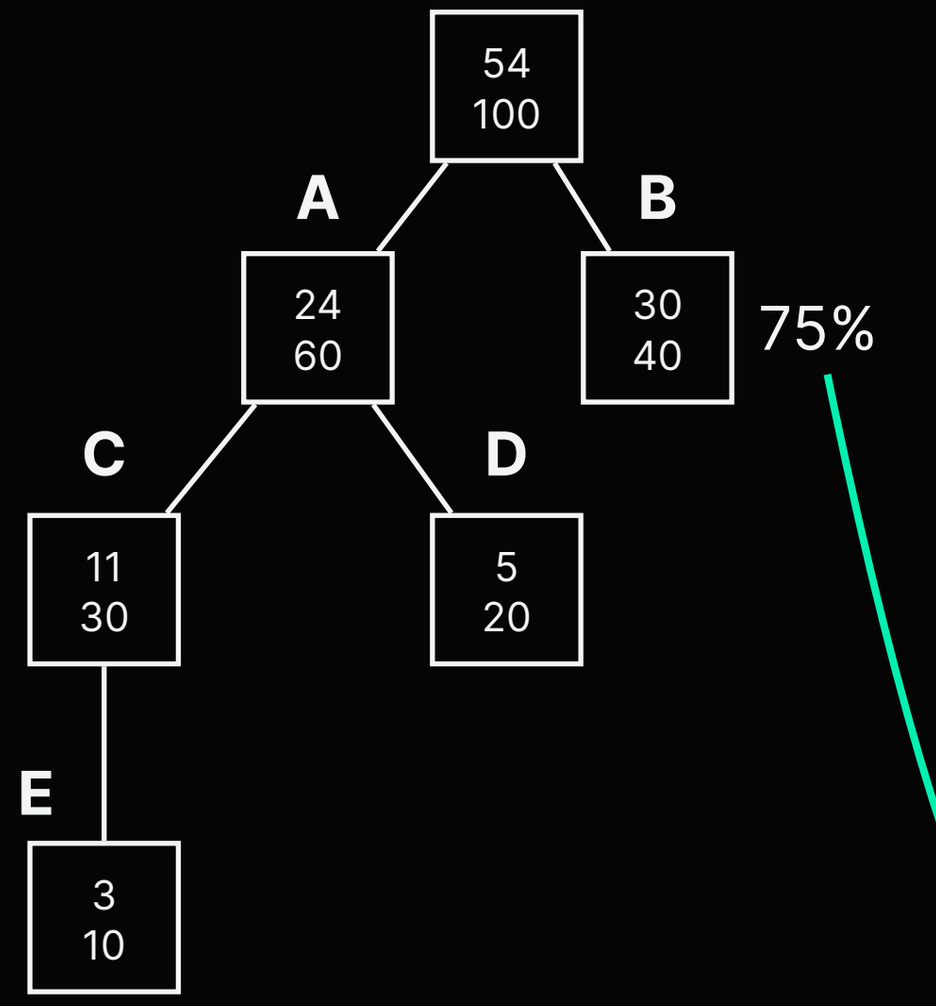
	A female	B male
START		
Sleeping	100%	100%
Sleeping		
Suffix	29.8%	29.3%
Washing		
	11.3% **+5.2%** 16.5%	
Commute		

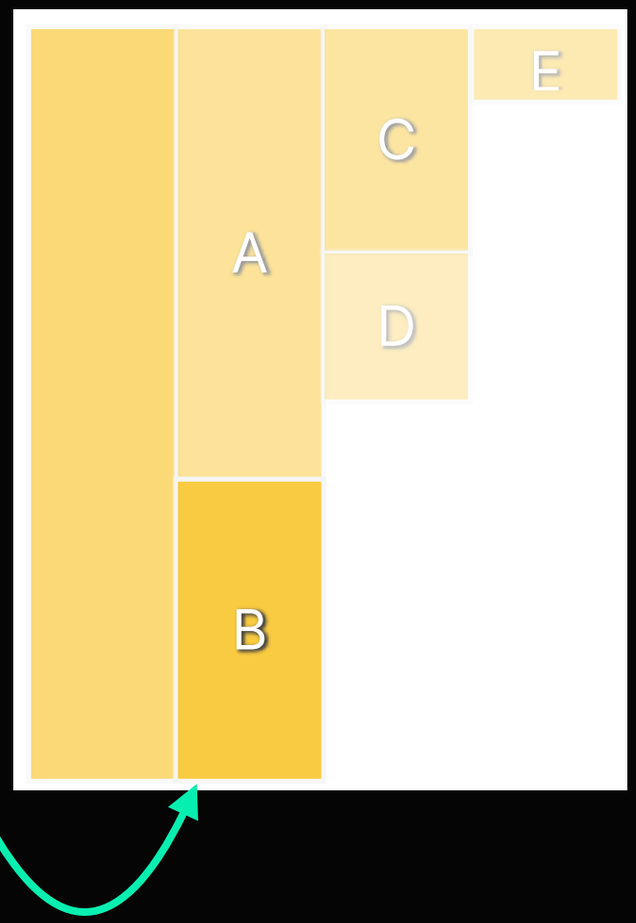
American Time Use Survey

P(Commute | Sleeping → Washing, Sex=X)

$\Pr(Y \mid S)$ when $Y \in \{0,1\}$

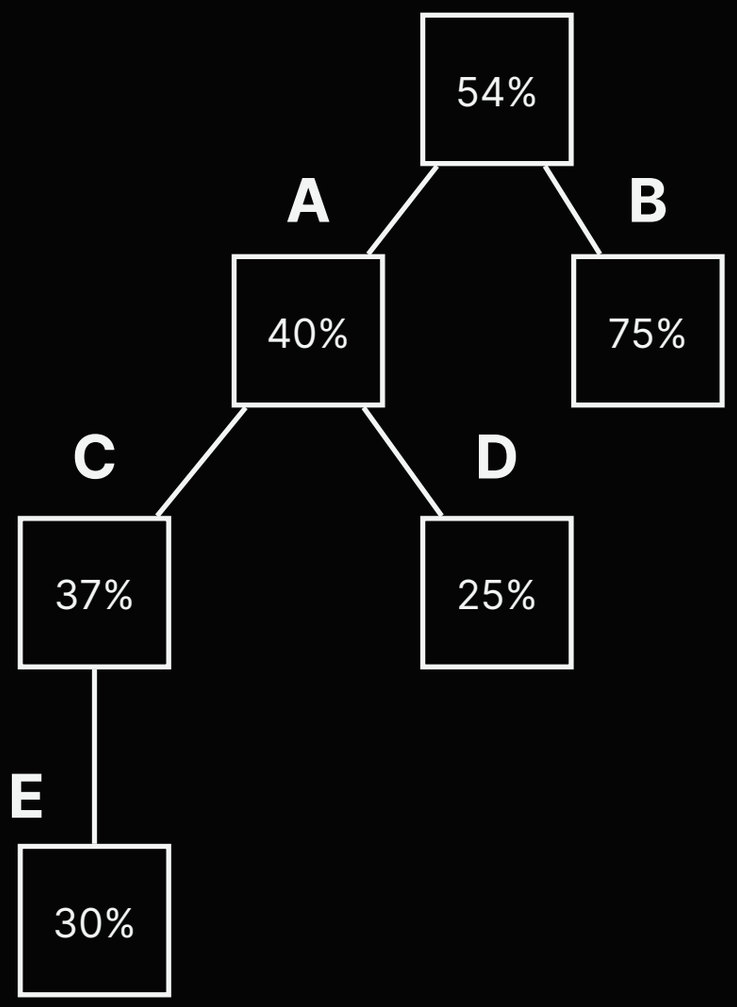
Sequence	Count	Successes
A	10	8
B	40	30
A→C	20	8
A→C→E	10	3
A→D	20	5

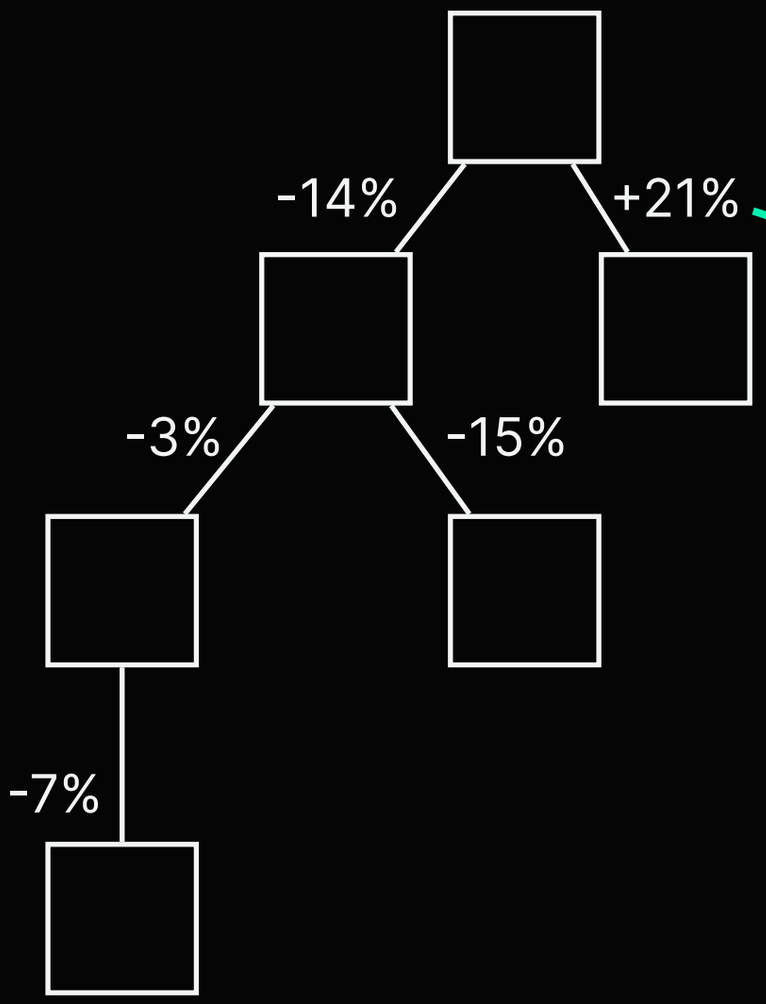


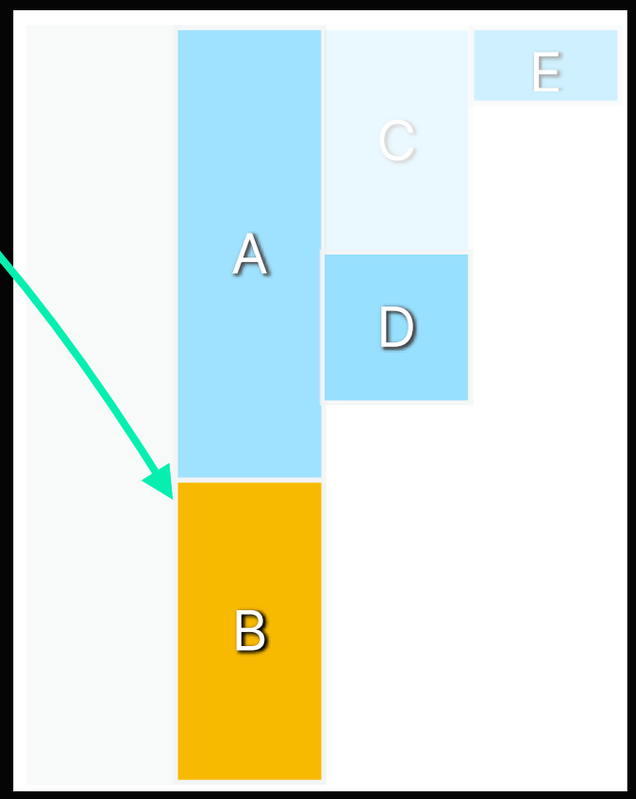


Raw data **Tree representation** **Visualization**

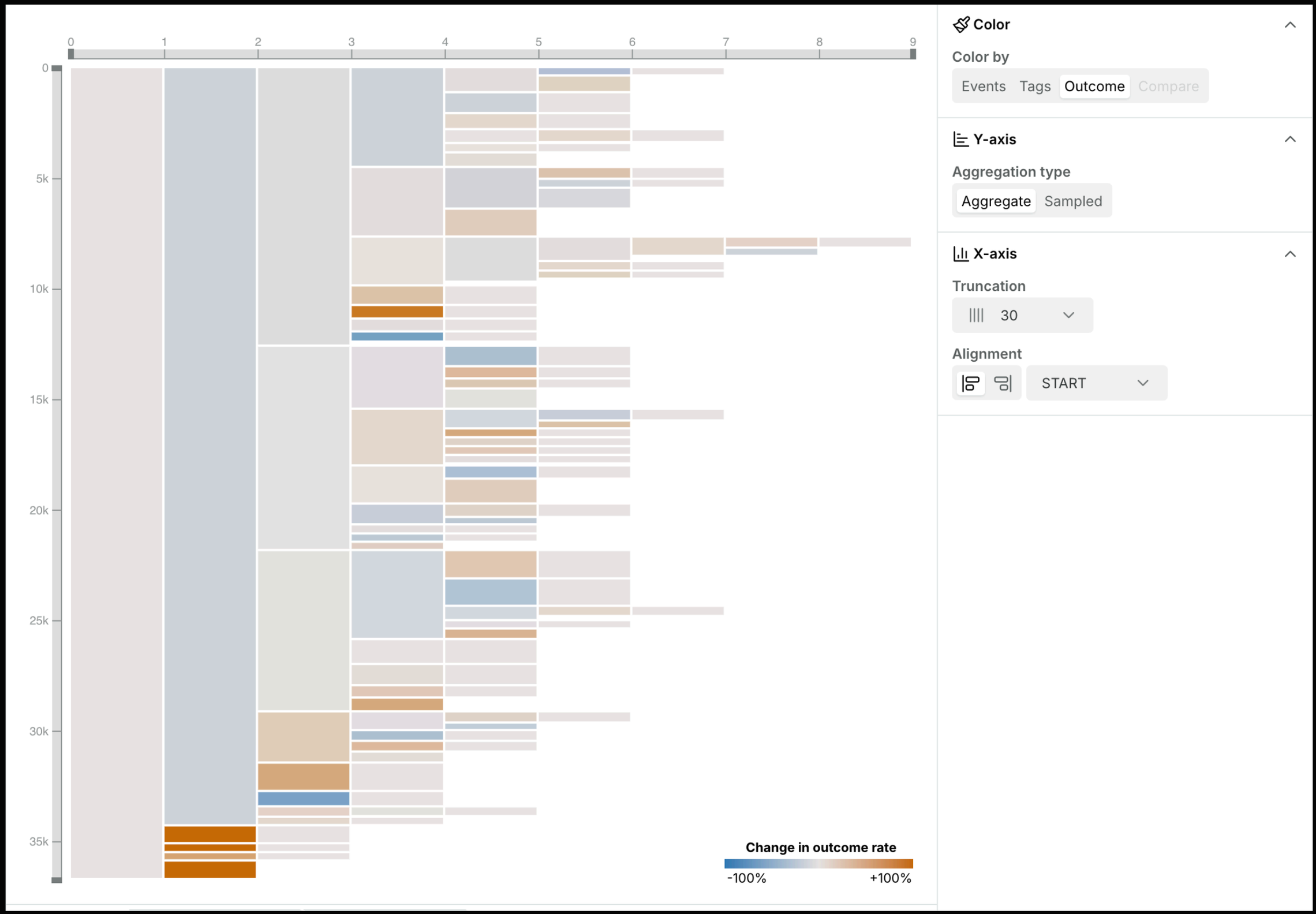
$$\Pr(Y \mid S_{1:k}) - \Pr(Y \mid S_{1:k-1})$$





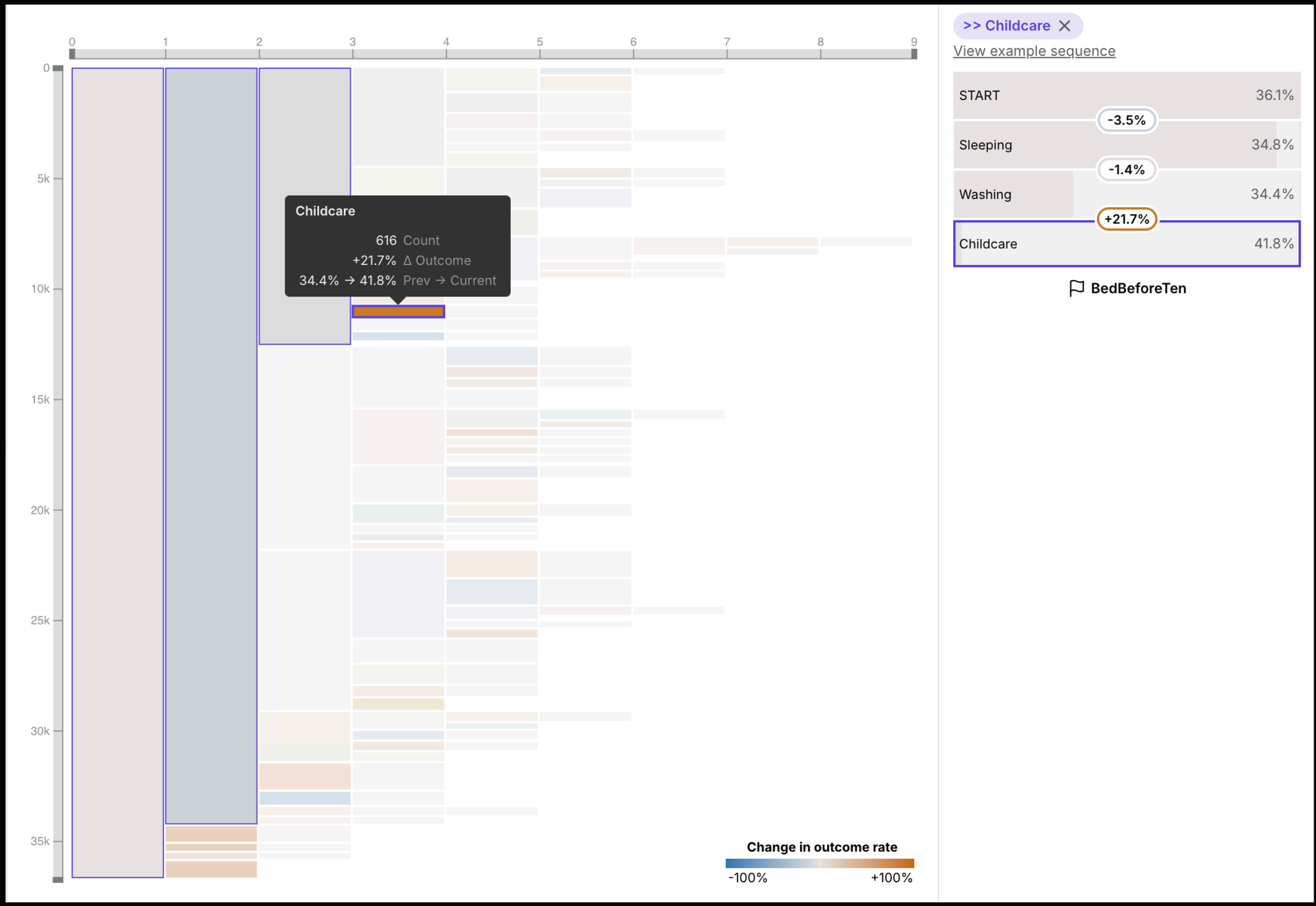


Node Rates | **Changes in Rates** | **Visualization**



American Time Use Survey

P(Early Bedtime | Sleeping → Washing → Childcare)



>> Childcare ×

View example sequence

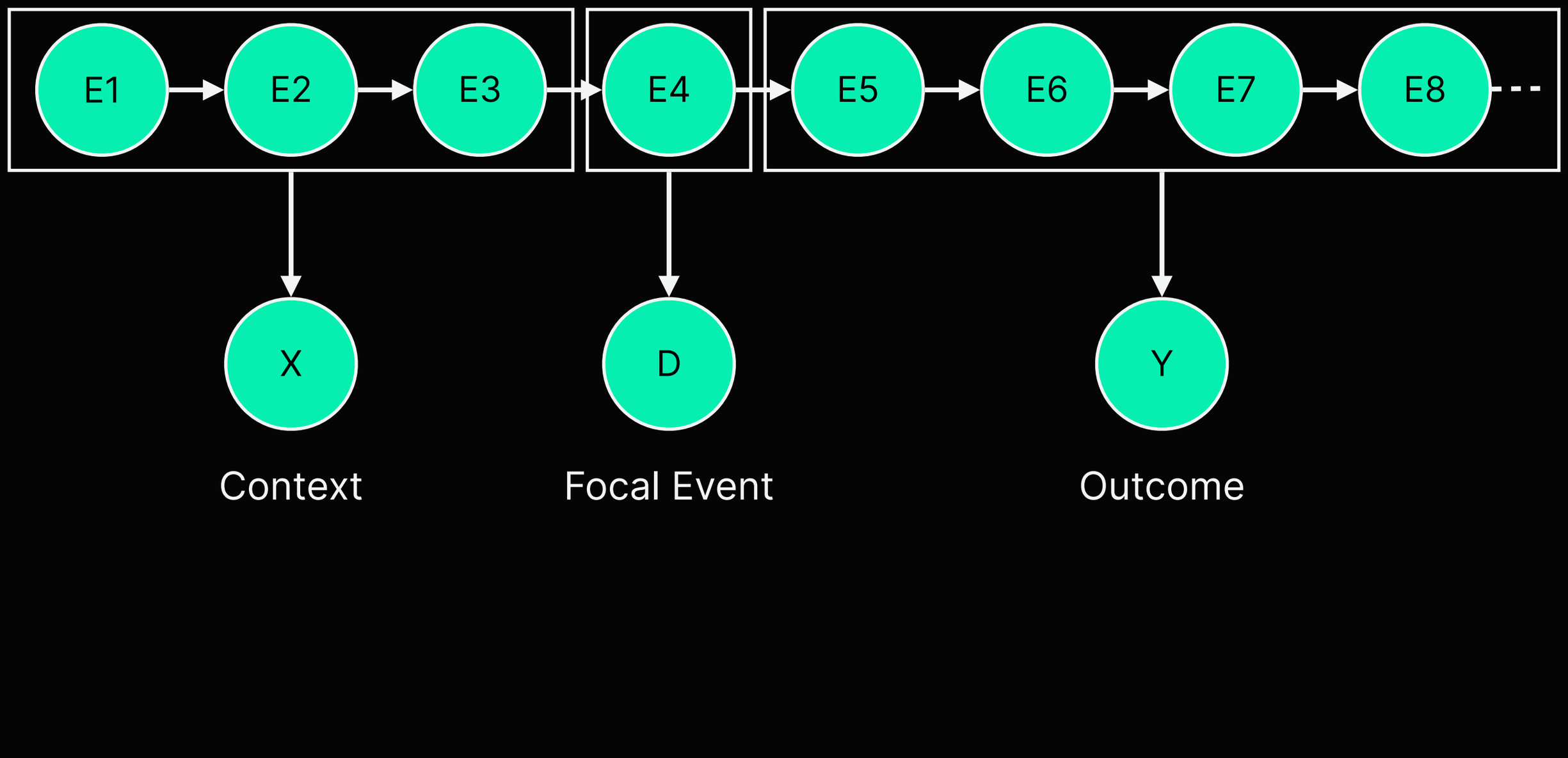
START	36.1%
-3.5%	
Sleeping	34.8%
-1.4%	
Washing	34.4%
+21.7%	
Childcare	41.8%

BedBeforeTen

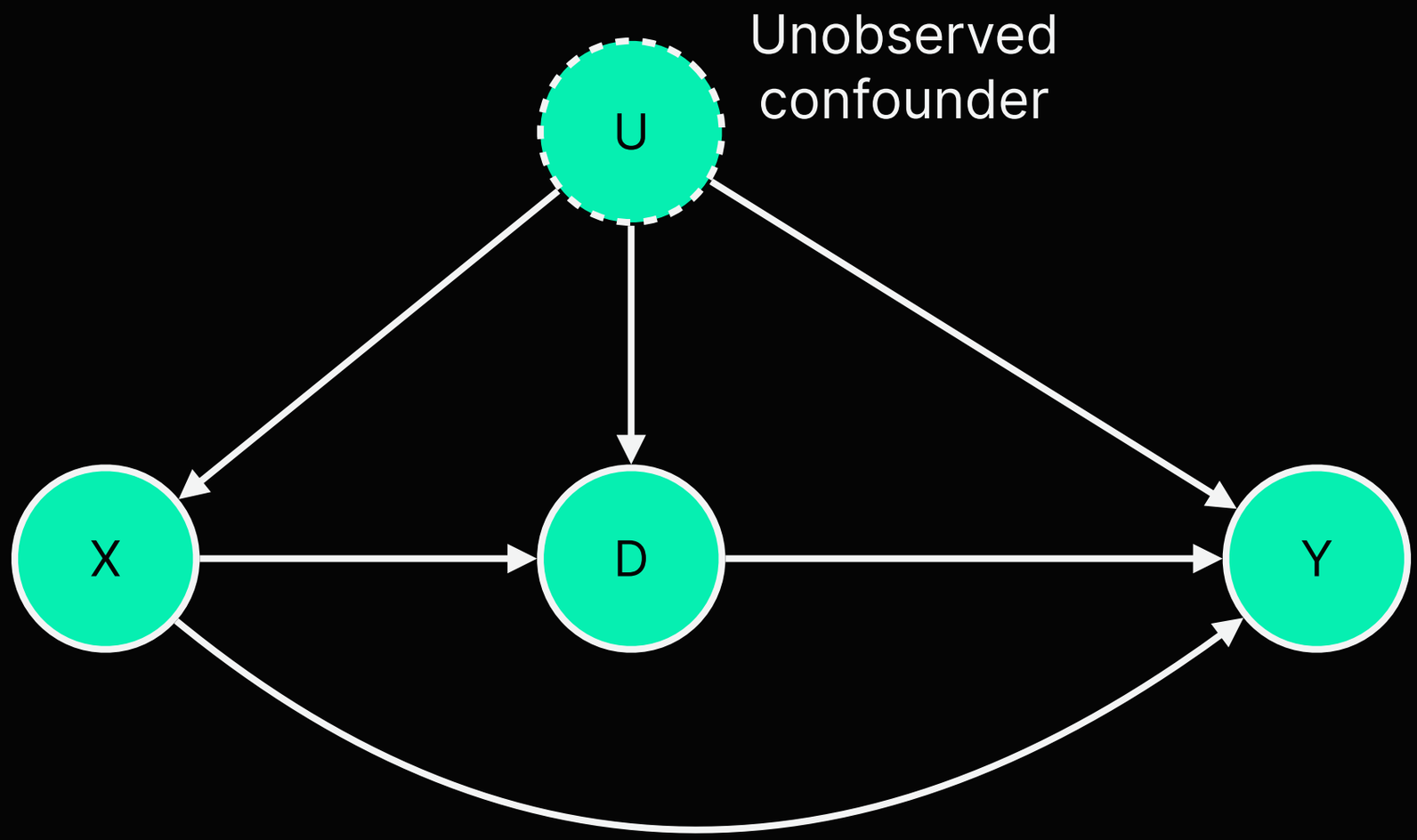
American Time Use Survey

P(Early Bedtime | Sleeping → Washing → Childcare)

Sequence to Variables

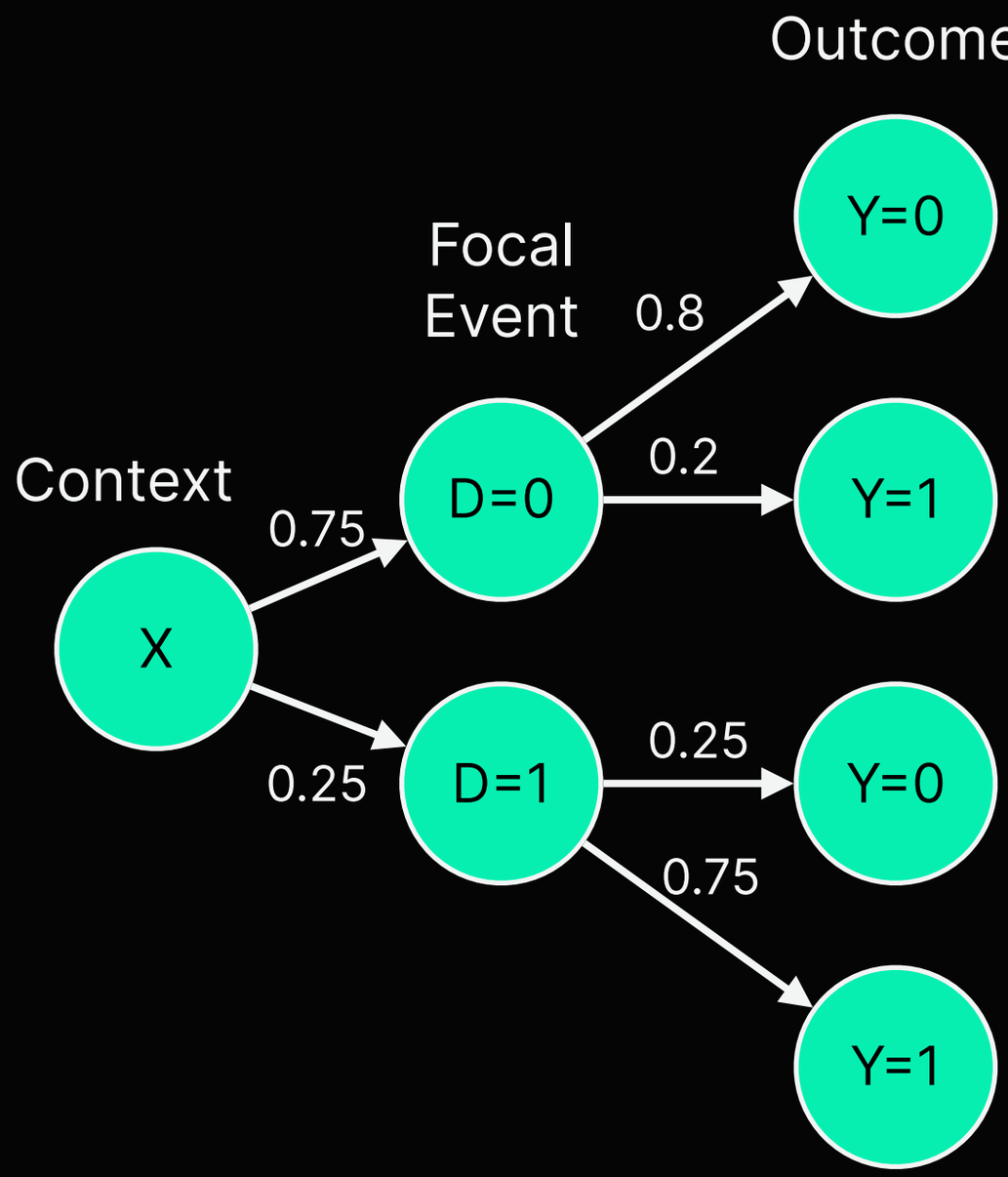


Sequence to DAG



What happens when we adjust for X?
(if only we had more time...)

“Interesting” Events



Context	D	Pr(D|X)	E[Y|D,X]
X	0	0.75	0.2
X	1	0.25	0.8

Key Questions:

- **Density**: How common is this situation? $\Pr(X)$
- **Control**: Can we change $\Pr(D|X)$?
- **Meaningful**: Is the difference large?
  $E[Y|D=1,X] - E[Y|D=0,X]$
- **Causal**: If we intervene will the difference persist.

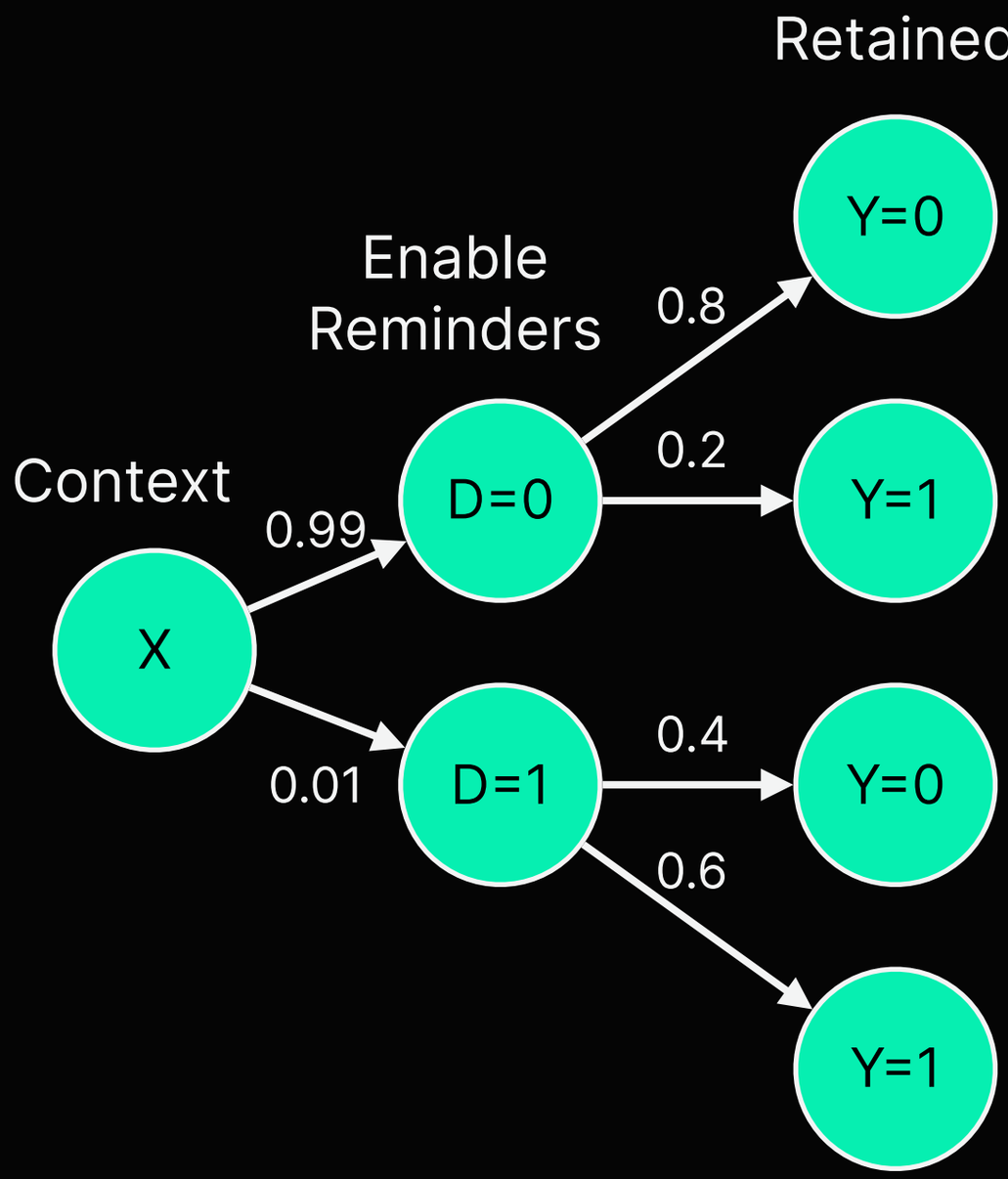
How Calm Increased Retention 3X After Switching to Amplitude

Discovering a Retention Hook

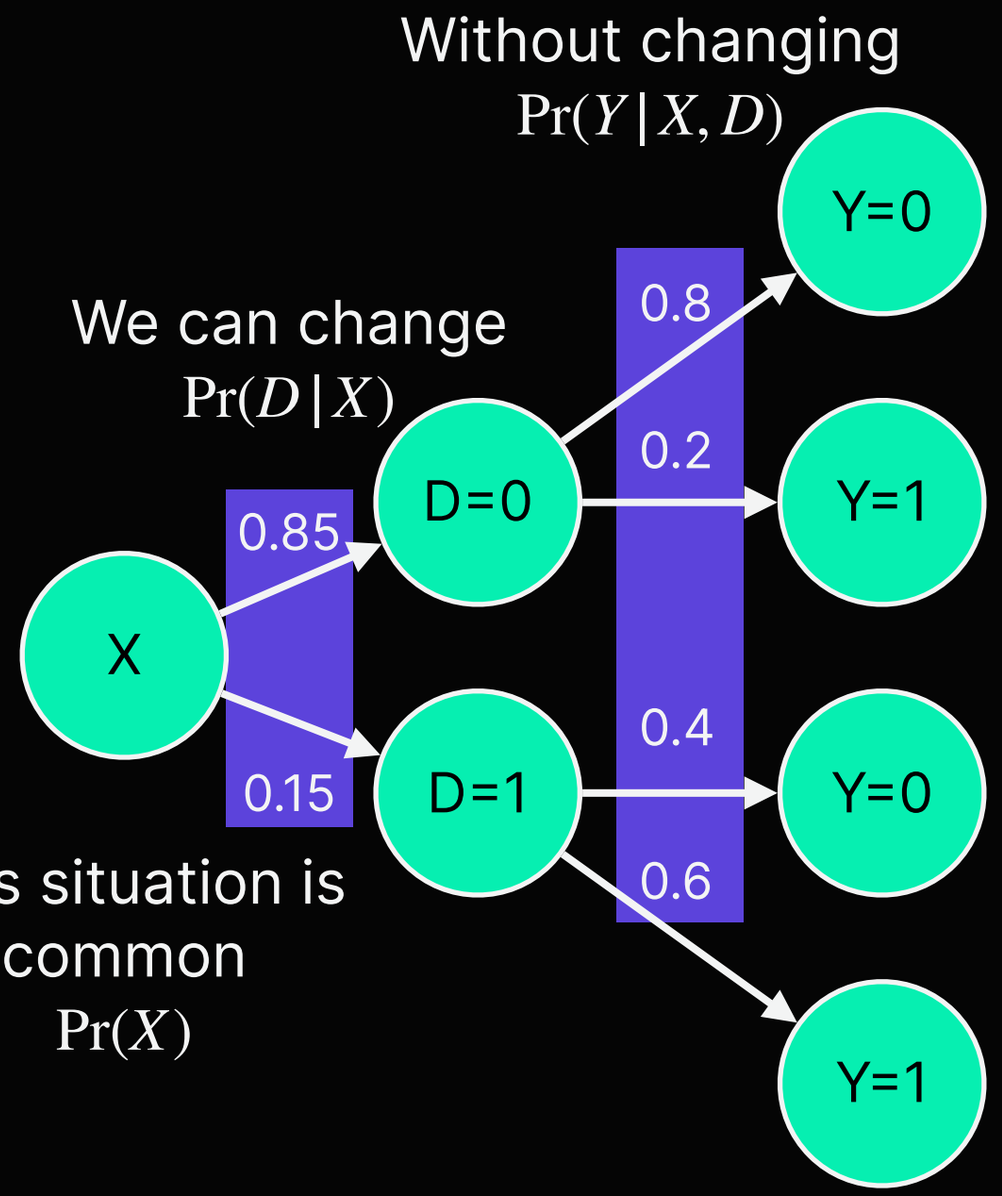
Calm built a Daily Reminder feature that allowed users to set a reminder for their daily meditation session, but the reminder feature was buried deep on the Settings page of the app. Very few users, less than 1%, were finding and using reminders. Calm hadn't given the reminders feature much thought for several months, but on a hunch, used Amplitude's behavioral cohorting feature to measure the retention of users who set a reminder, compared to those who didn't.

To their surprise, they found an almost 3x increase in retention for users who set Daily Reminders. With such a small sample size of users, they couldn't know whether this was a causal relationship. It could be that the power users of their app, who would have been well-retained anyways, were the ones digging into the Settings page and finding the Reminders feature.

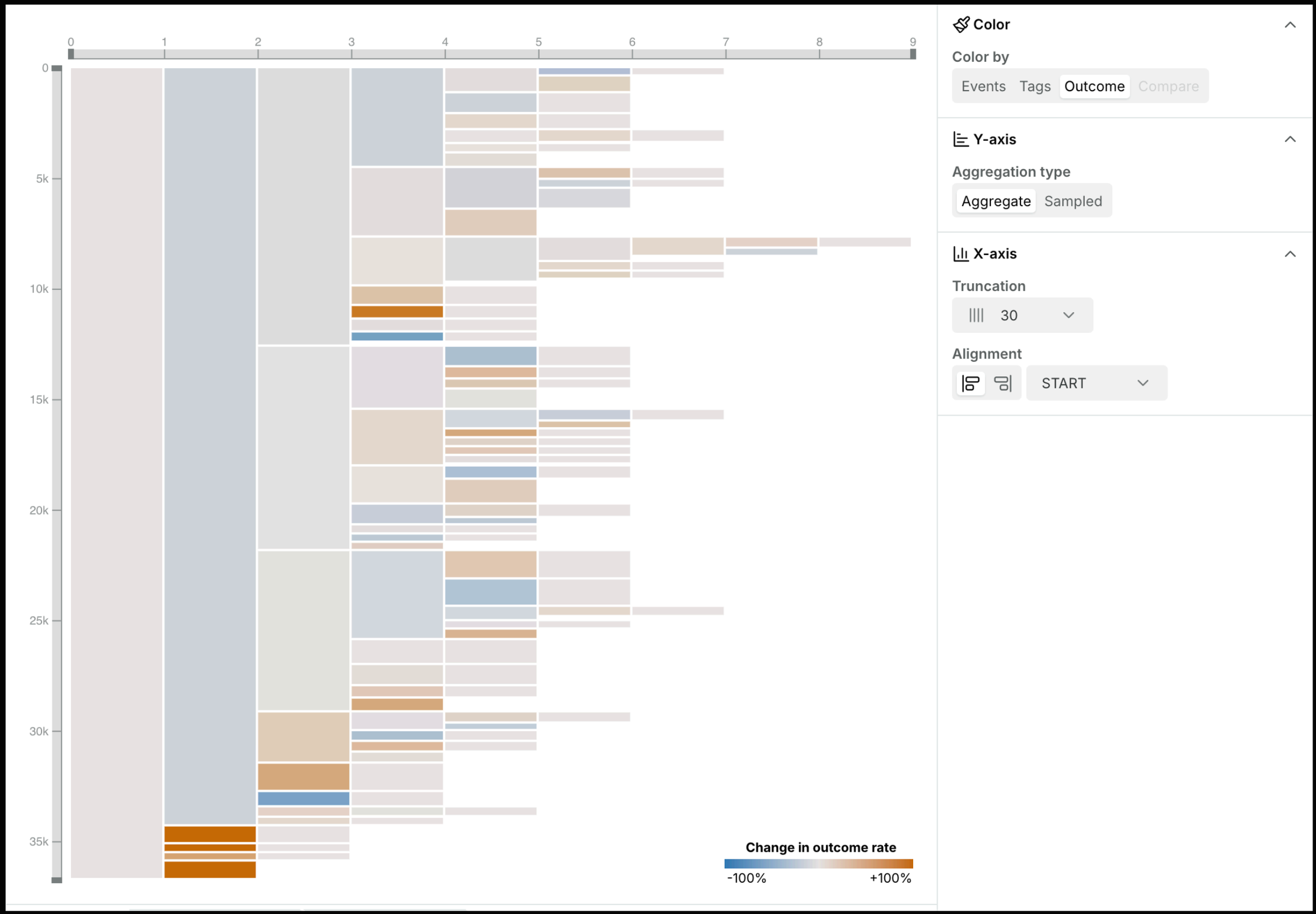
Hypothetical Intervention



Factual



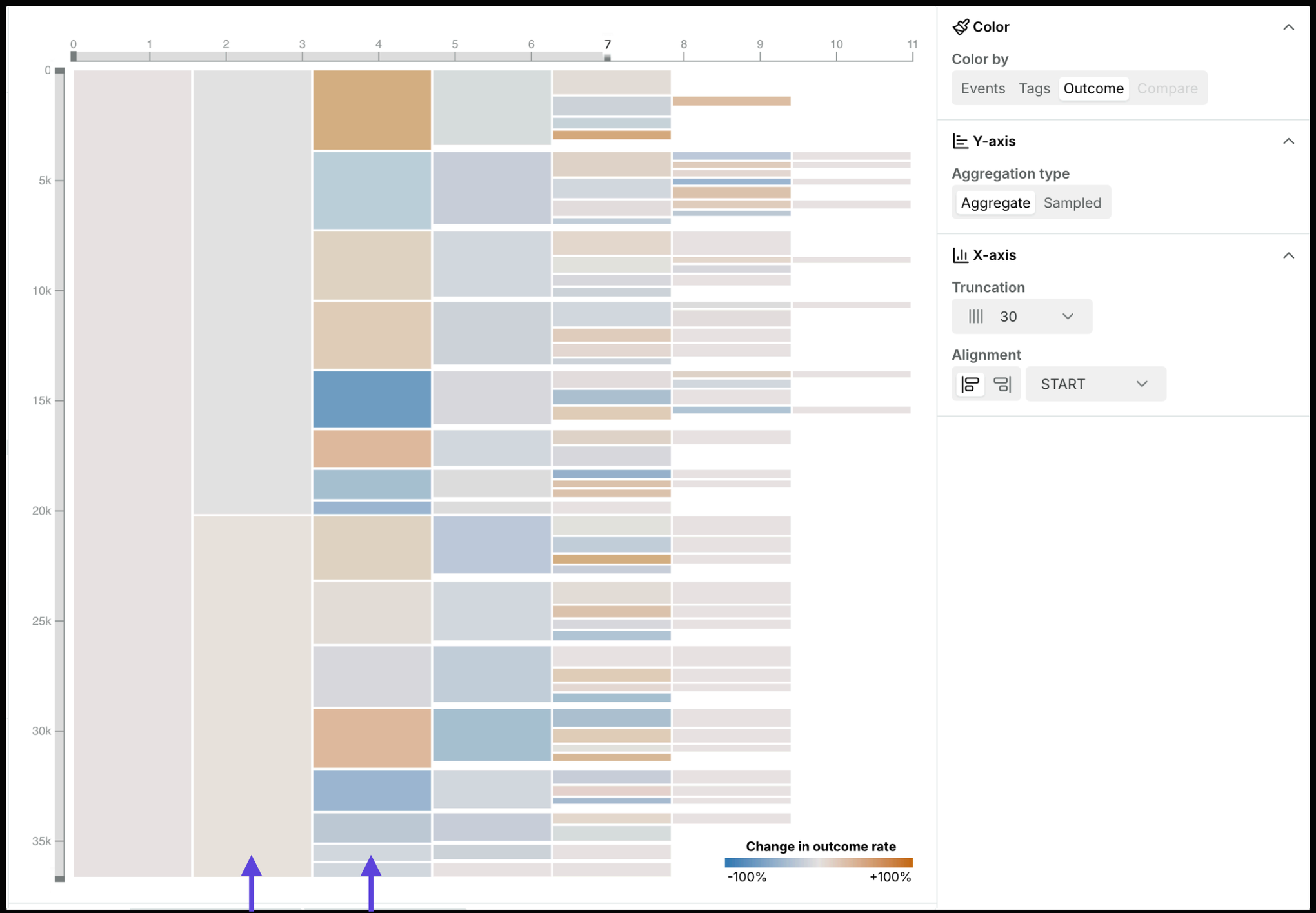
Counterfactual



American Time Use Survey

P(Early Bedtime | Sleeping → Washing → Childcare)

Adjusting for pre-sequence information

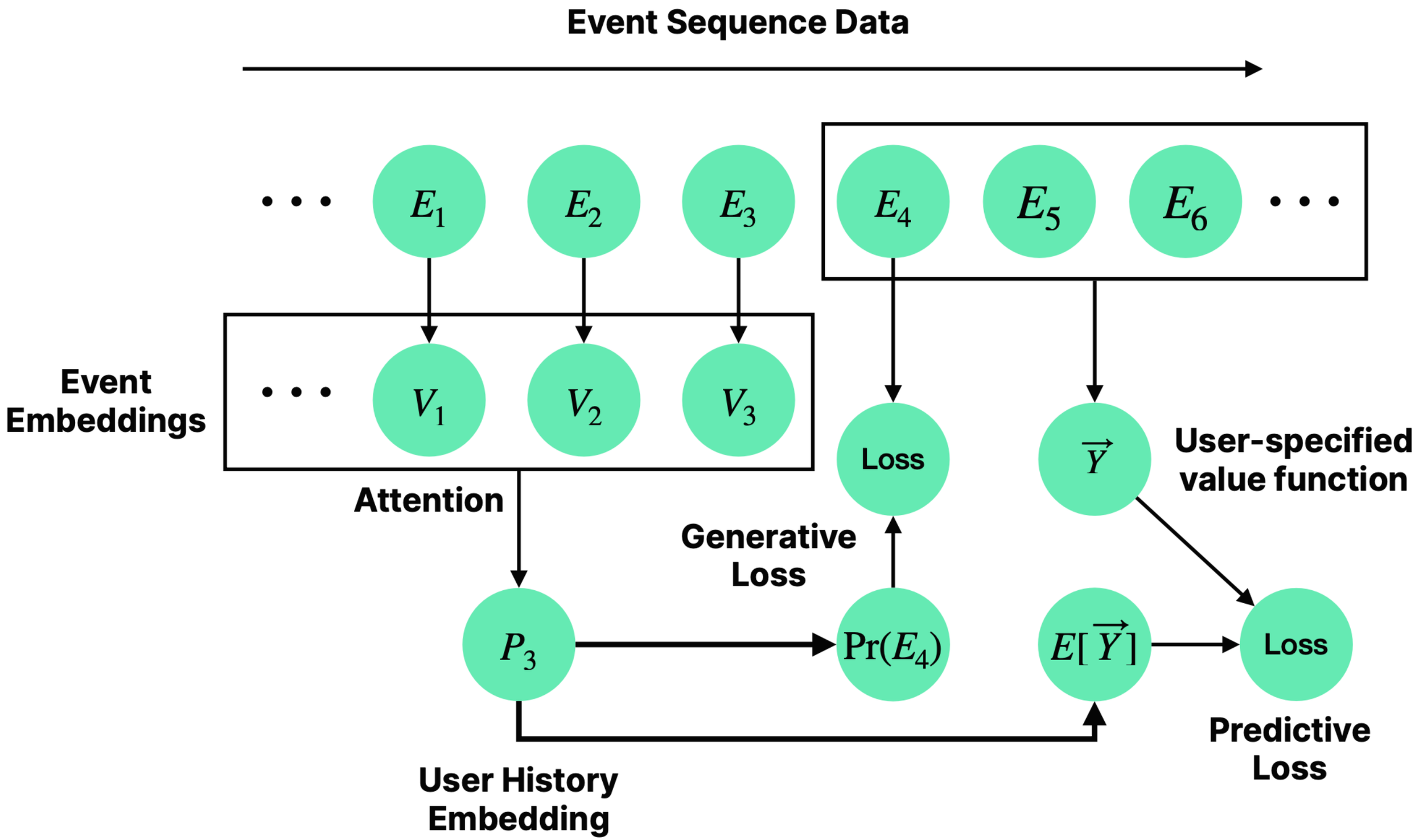


Sex Age

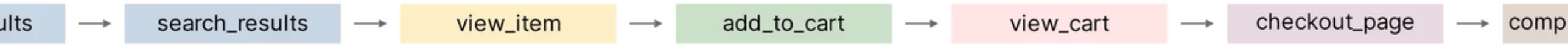
American Time Use Survey

P(Early Bedtime | Sex, Age, Activities)

The Future of Encoding the Past

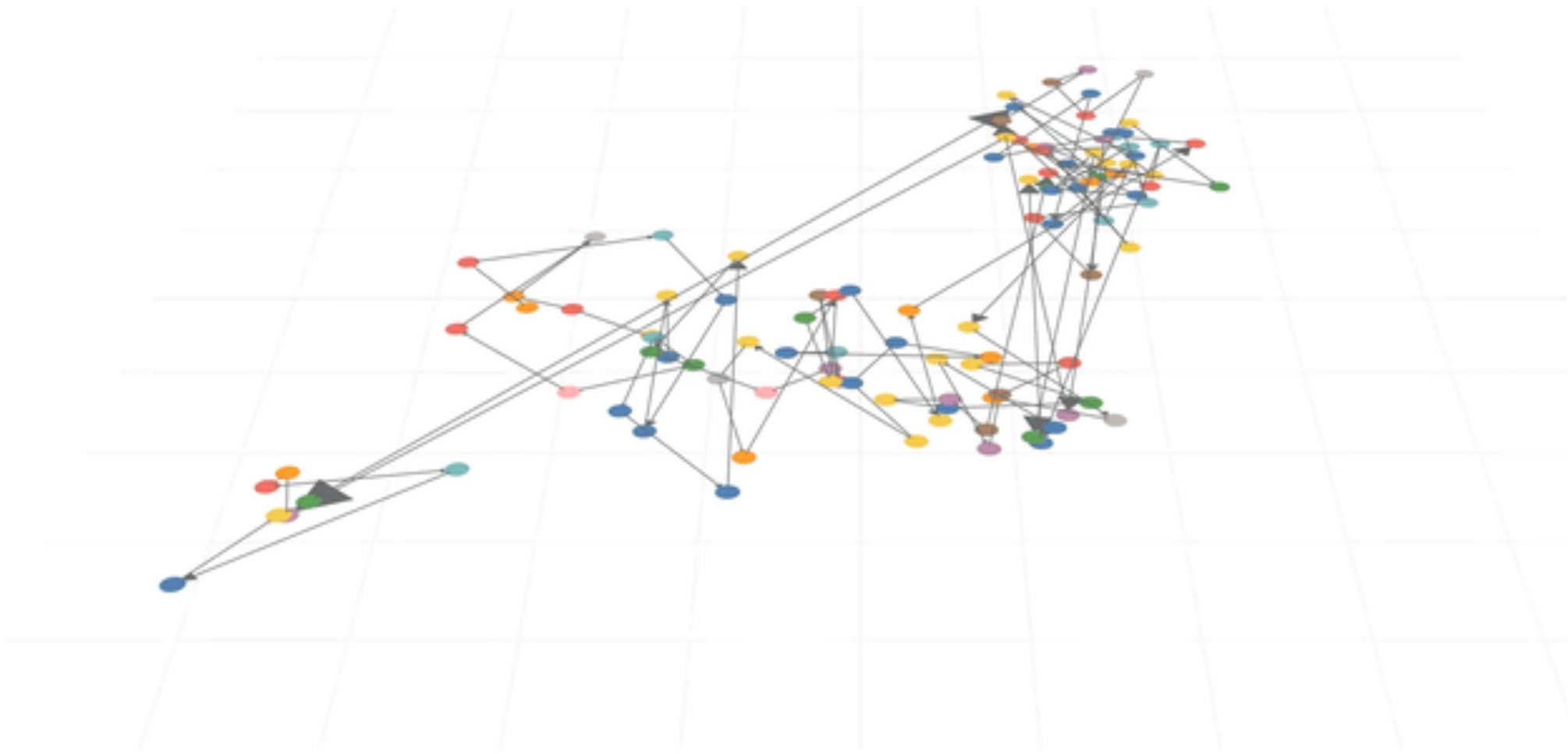


Original sequence



estimates.

Latent space embedding



Summary

- Sequence data is very general and very common.
- It remains challenging to work with — we need new tools and approaches! Can leverage rich visualizations with interactivity.
- When modeled and visualized properly, it can help with advanced analysis tasks aimed at generating new hypotheses:
    - What is different between two sets of sequences? (e.g. a launch or product test)
    - Which events contributed meaningfully to an outcome of interest?
- One of the modeling challenges is efficiently encoding an event history, and advancements from LLMs provide a promising approach.