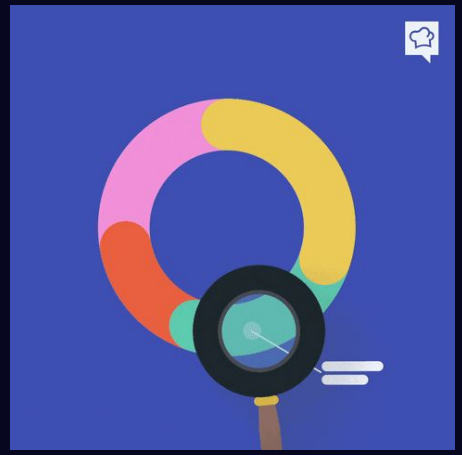
Nine
PRODUCT/
TECHNOLOGY



Data Analytics Competition

31 October 2020

Flashback to the skills workshop

We mentioned that *Cookies* are basically dead now. Local and Session Storage is the modern equivalent.

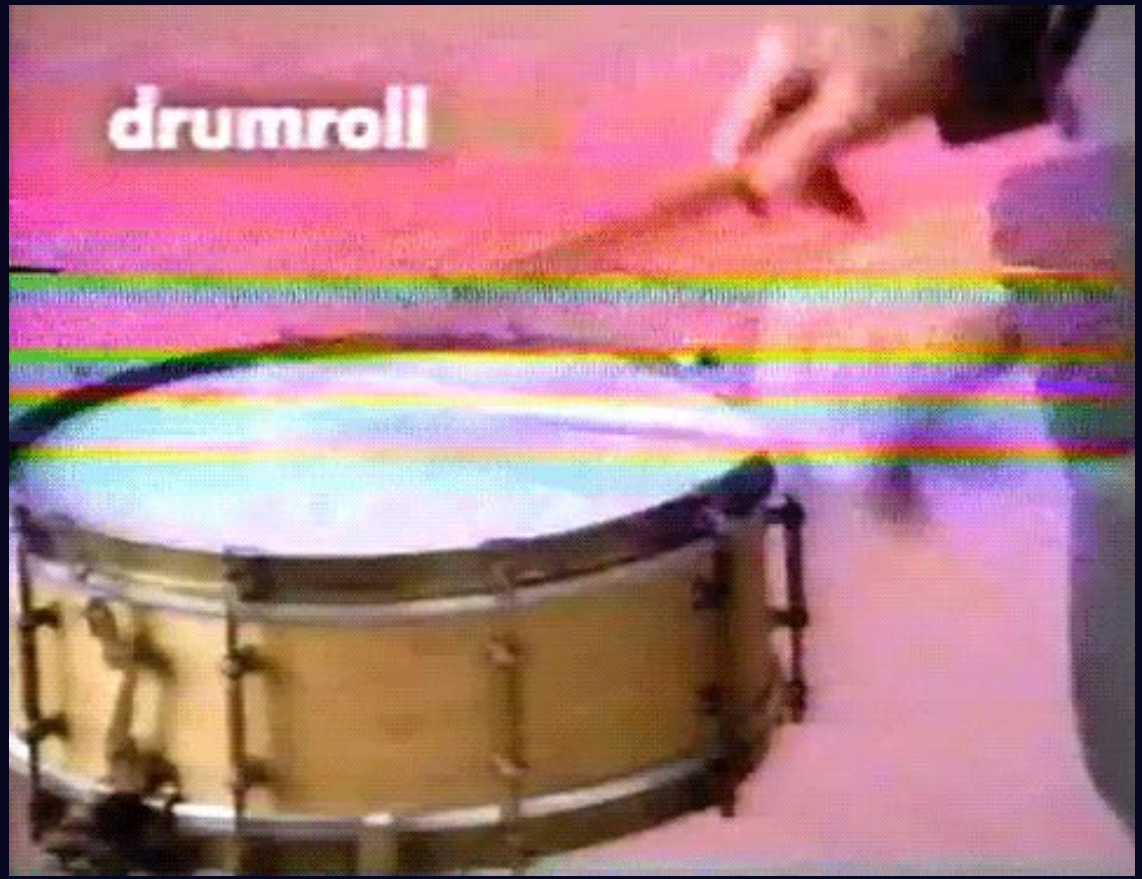
- Cookies/Local Storage is a simple browser attached datastore that can be used to cache **application state**
- **Local storage** persists until cleared by the user or by incognito/private browsing ending.
- **Session storage** persists until the browser is closed.

Flashback to the skills workshop

- When a browser visits our domain/site for the first time, create a UUID and store it in **LocalStorage** we'll call that our **ClientID** (or BrowserID). This persists until the browser cache is cleared.
- When a browser visits our domain/site, create a UUID and store it in **LocalStorage with a timeout of 30 minutes** we'll call that our **SessionID**. Whenever there is any site activity extend the timeout by 30 minutes. This persists until the timeout is reached, midnight occurs or browser cache is cleared.
- When a browser requests a page create a UUID and store it as a variable on in the page. We'll call this our **PageViewID** and it persists until another page is requested.

So, why do we care about accurate data?

- We want to deliver world-class content for our audience. One of the most powerful product development tools we have is reporting on how different features of the site are being used.
- But what happens when you throw in **non-human traffic** or audiences that are sharing accounts? This data can **skew our ability to make the best decisions** for features and content that our readers really want.
- Not all site visits are equal!


drumroll

Create an application that can detect **unusual levels of site activity**

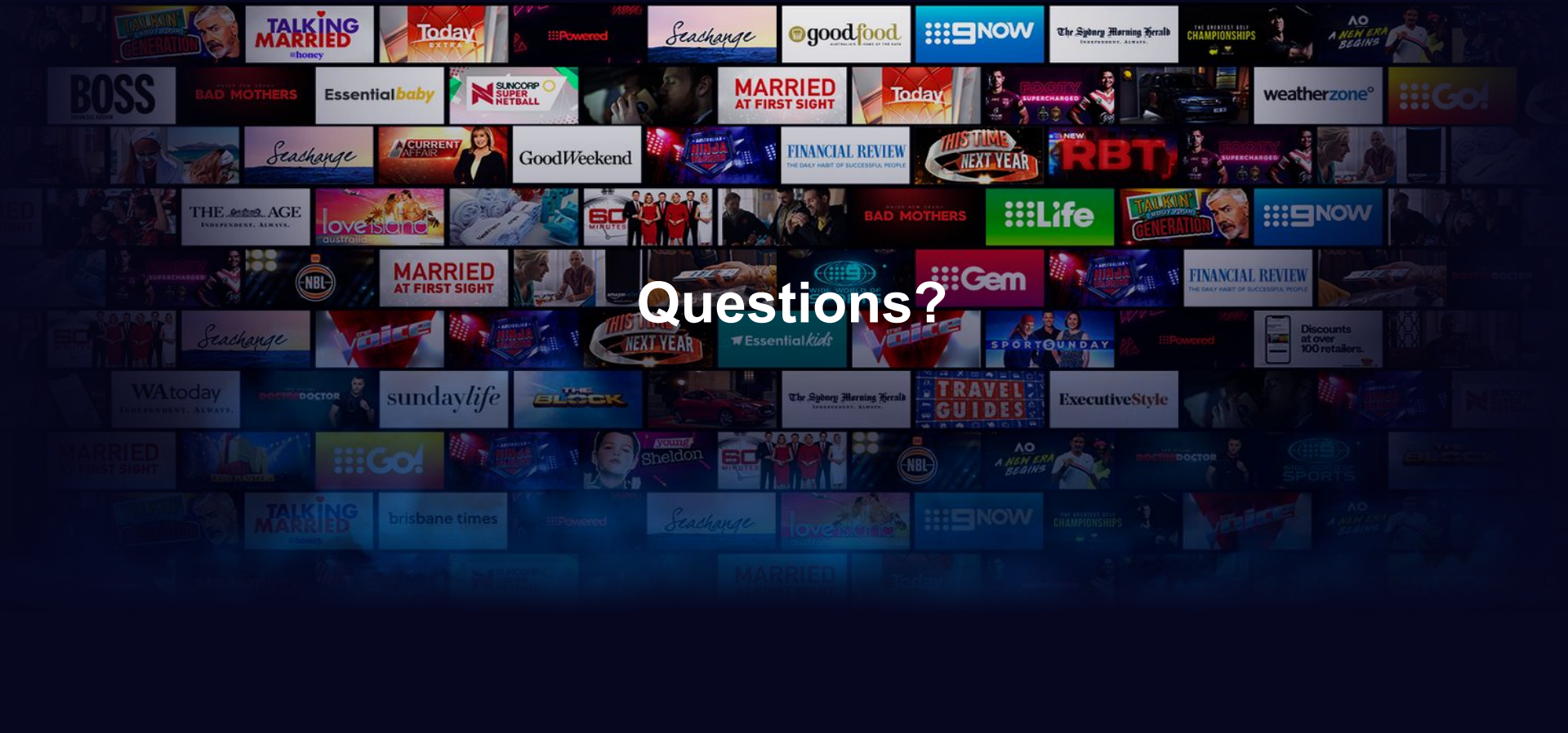
- Bots don't always declare themselves. Can you identify them?
- Can you identify indicators of user accounts that have their identity stolen or their accounts shared?
- Are there network architectures which are skewing our audience numbers? For example, Proxy services?

Why is this important?

- Bot traffic can be malicious - identify and block DDOS attacks
- Bot traffic can create outlier data - is that a poor performing story or a bot visit?
- Advertisers do not want to pay for bot traffic on their ads - IVT rates.
- We want to identify scrapers that are stealing our content to republish without a license
- We want to identify subscribers that are sharing their logins or have had their login stolen

The Ask

1. Can you take a period of historical data and build a detector that flags unusual user behavior or Impossible Subscribers?
2. Can you build a classifier that can suggest what type of user they are?
3. Can you detect bots that declare themselves versus bots that are being sneaky?
4. Can you see potential new subscription product offerings emerge from the data?


Questions?


Nine
PRODUCT /
TECHNOLOGY