Digital Commons™

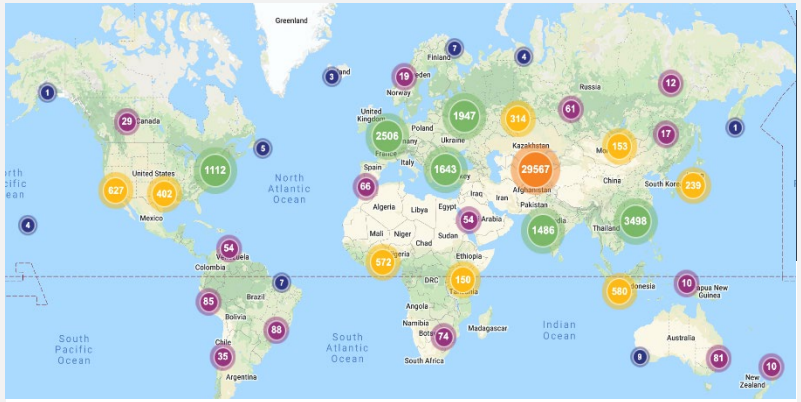
Fighting the Bots: An Update on Digital Commons Download Filtering

Jiaqi Liu
Digital Commons | Elsevier

Riding the Wave

Digital Commons North American Conference 2021
October 26th-28th

What happened?



Readership Dashboard

As a marketing tool for the institution, the modern IR provides clean, accurate usage data to reliably indicate the reach and impact of the institution's content to stakeholders.

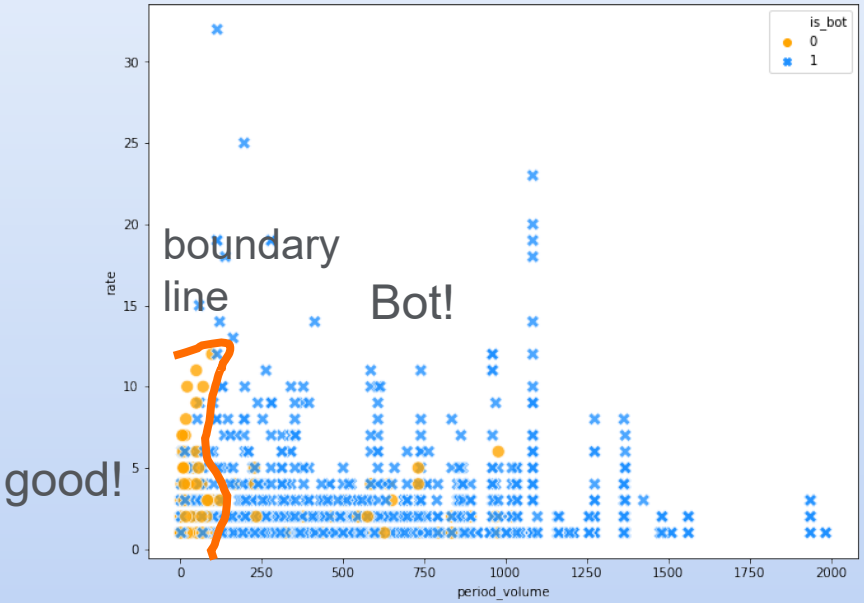
However,…

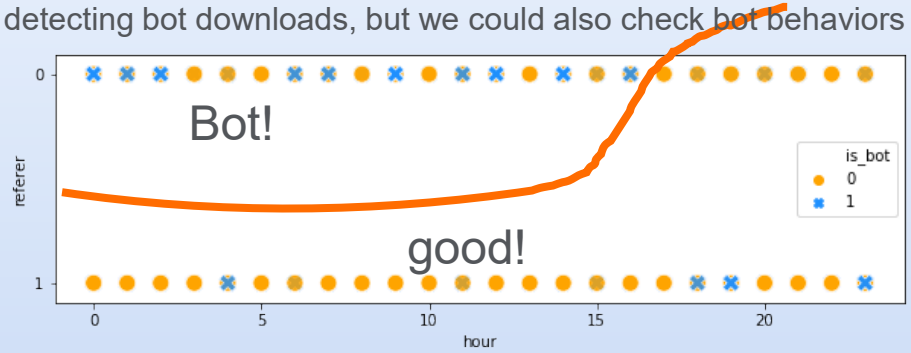
In a few very rare instances, we are observing artificially geo-distributed proxy farms of bots downloading IR content.

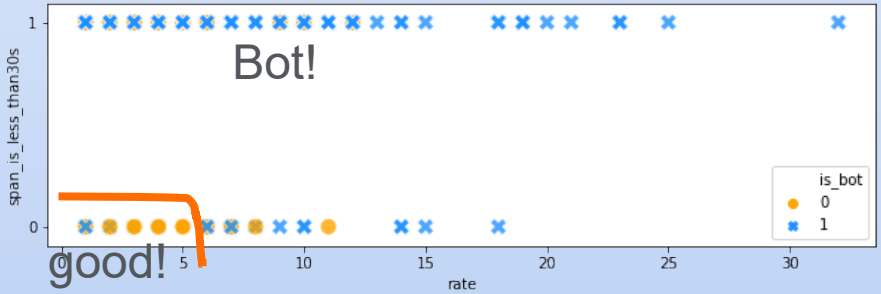
Detecting the bots

User clicks → Download logged → Download filtered → Download recorded in database → Send data to dashboard

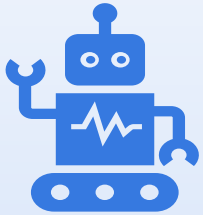
Currently, our IP-based filter system has sensitive of 67% on detecting bot downloads, but we could also check bot behaviors







Build a bot download filter



Bots Behavior

- Download rate
- Download volume
- Timespan
- Referrer
- Download percentage
- Cookies
- User agent
- ……

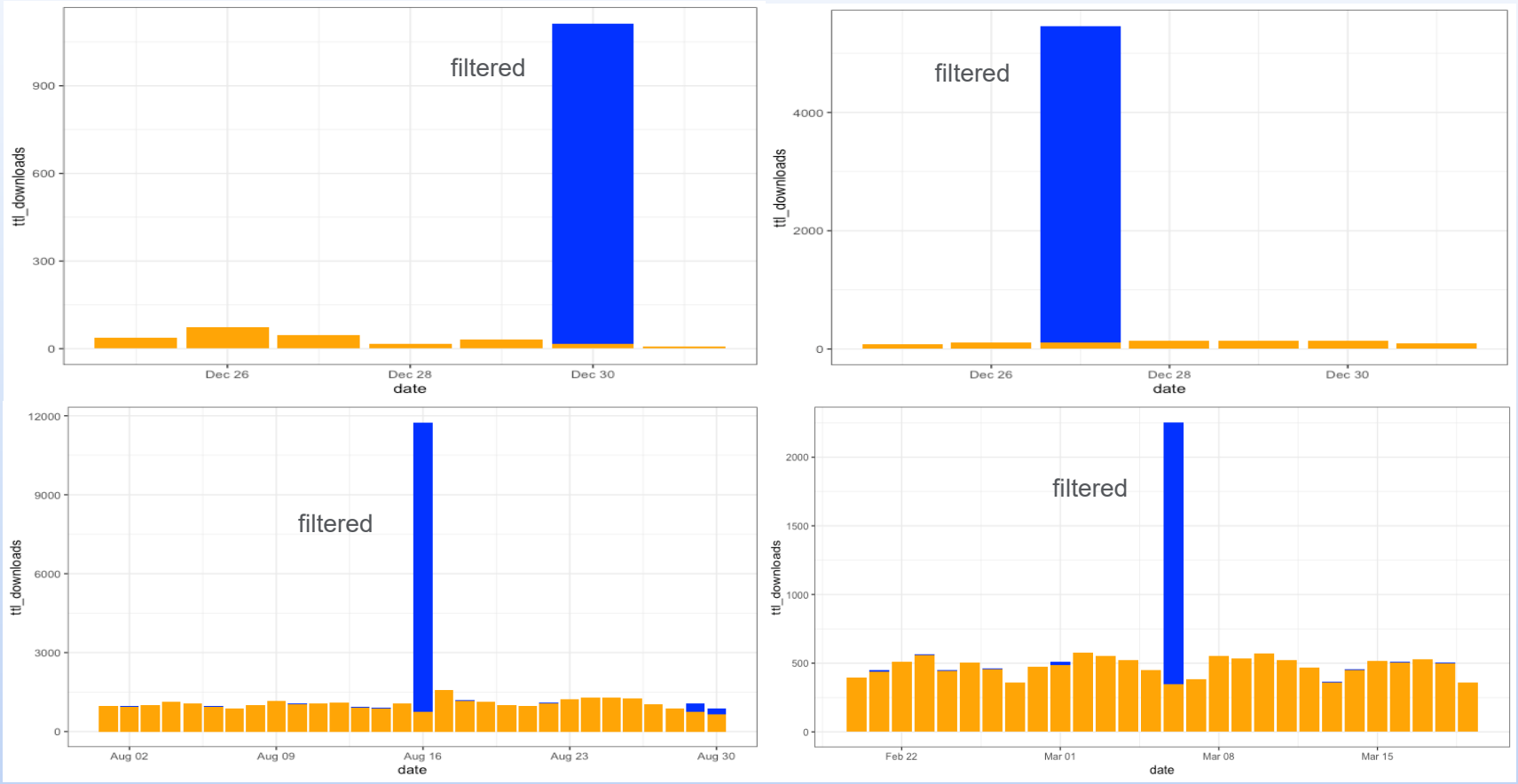
Download filter

- Created multiple boundary lines to segment legitimate and bot downloads
- Label the bot downloads as ’bot’
- Filter out bot downloads

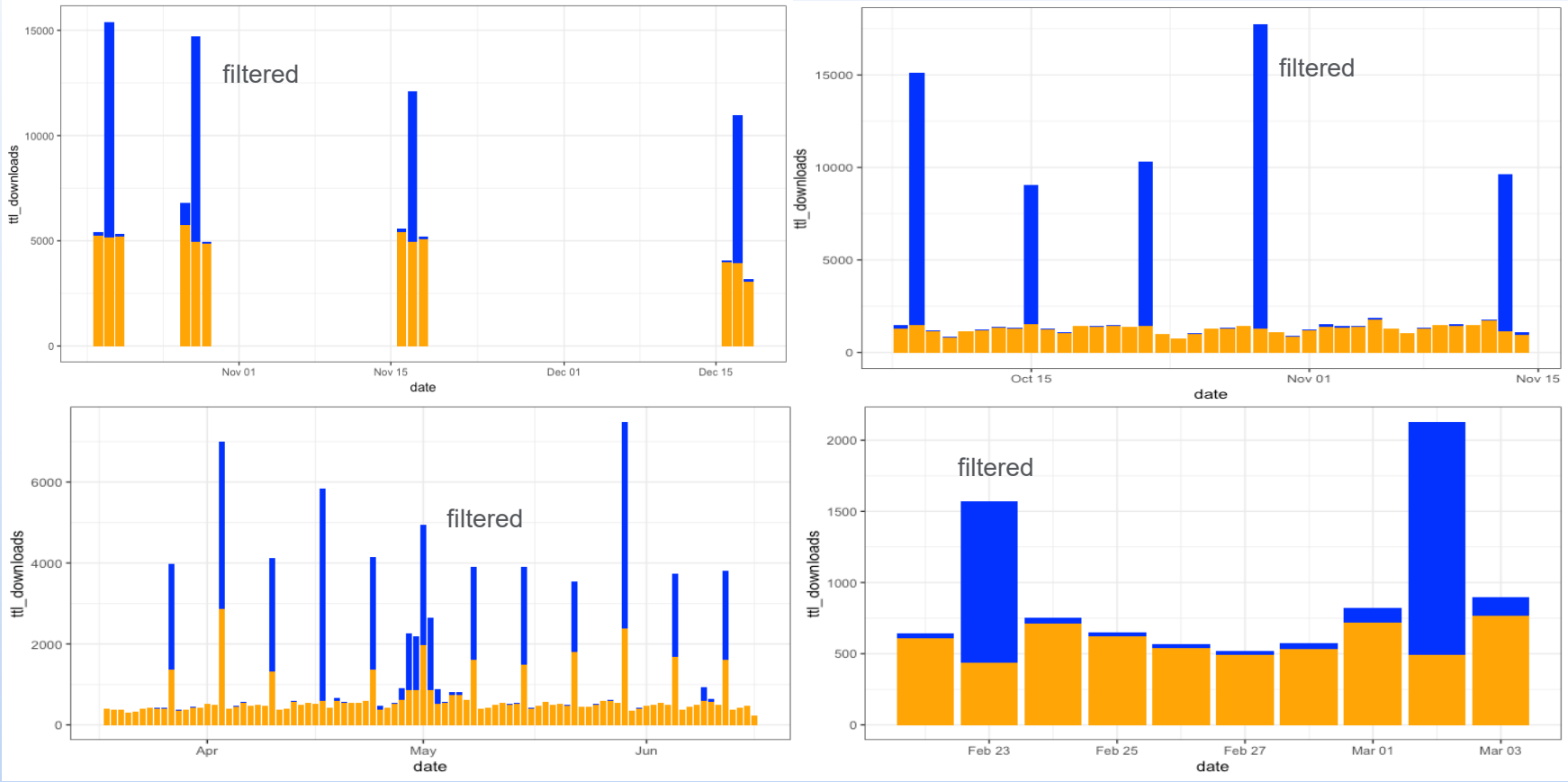
Legitimate downloads

- Publish only legitimate downloads to dashboards

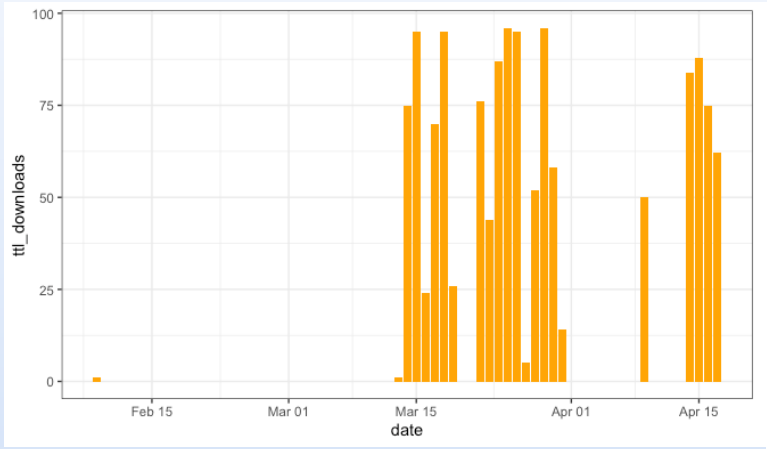
With a newly designed algorithm, single institutional spikes are removed!

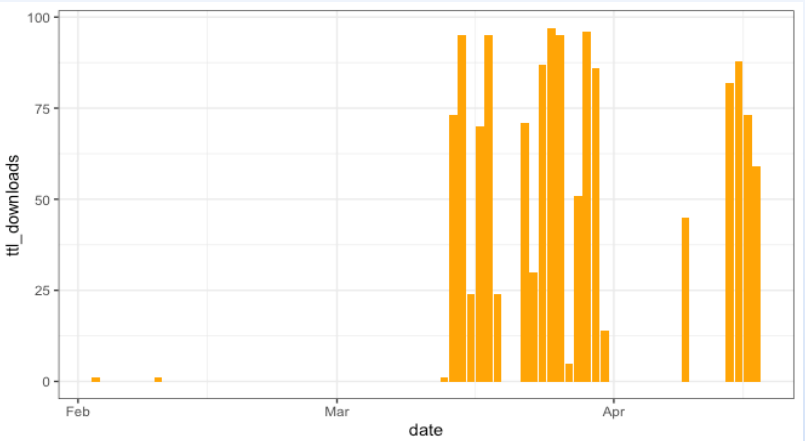


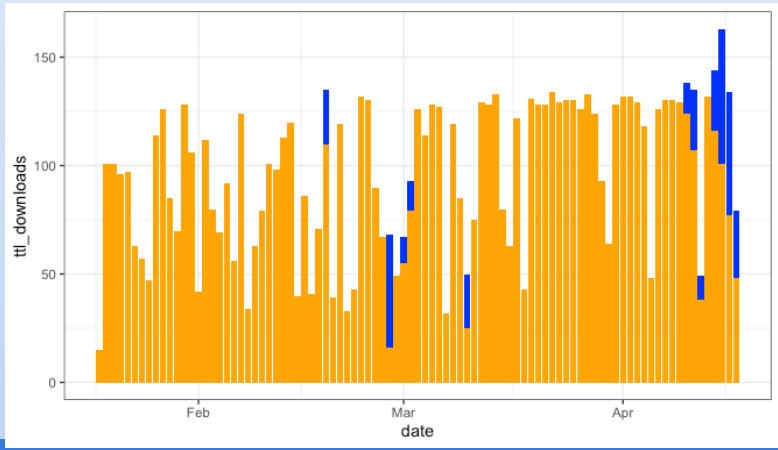
Periodic institutional content spikes are removed!

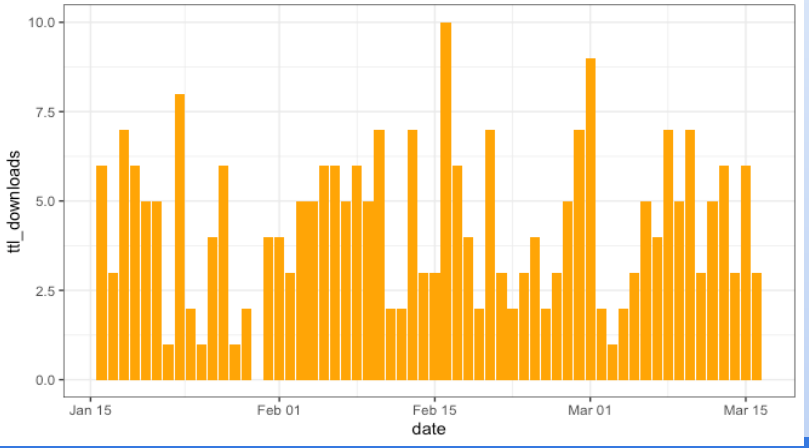


Sustained work-level plateaus are not removed; we are still working on it!

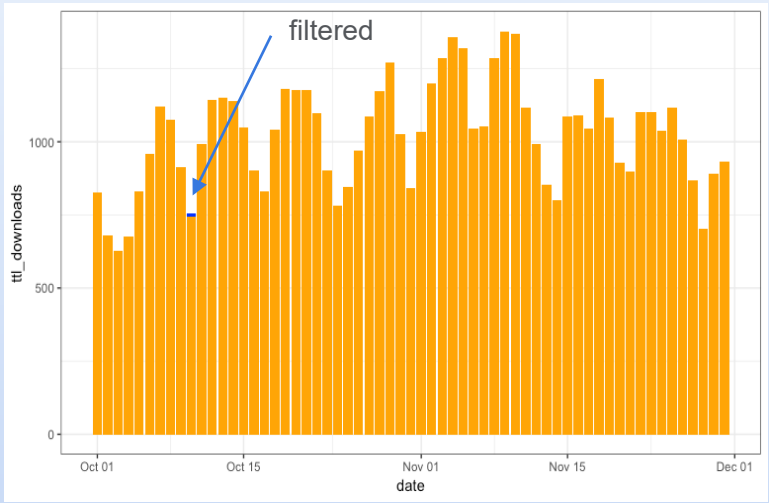


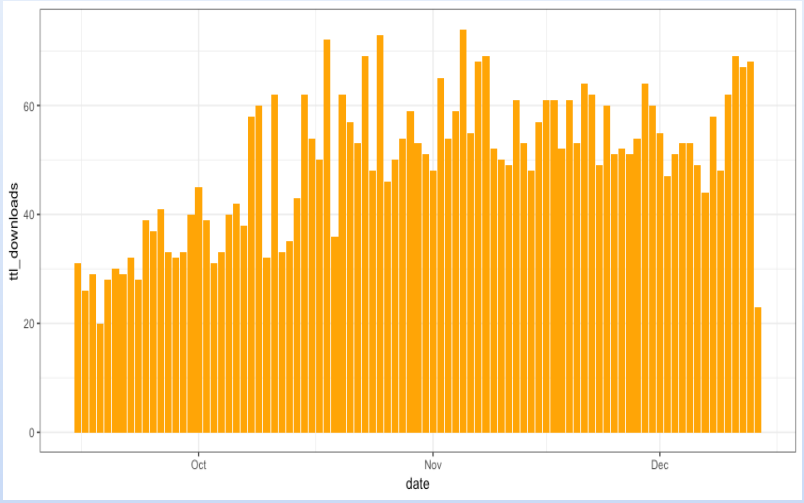






The new algorithm would not significantly affect legitimate downloads





Key takeaways

- Bot downloads have suspicious behaviors that distinguish them from legitimate ones.
- We've improved our abilities to identify and remove suspicious downloads, this will soon be reflected in your dashboards.
- We've ensured that maximum legitimate download will not be removed by the model to provide you with the most comprehensive view of your readership.
- We already have had the industry-leading download filtering capability, we'll continue to improve our bot detection model to provide better solution which will keep up with developing bot techniques as a new challenge.

Digital Commons™

Thank you

Jiaqi Liu
Bepress | Elsevier

Riding the Wave
Digital Commons North American Conference 2021
October 26th-28th