Participatory Web Archiving

Opening the Black Box of Save Page Now

JESSICA OGDEN
University of Southampton
jessica.ogden@soton.ac.uk
@jessogden

ED SUMMERS
University of Maryland
edsu@umd.edu
@edsu

SHAWN WALKER
Arizona State University
shawn.w@asu.edu
@walkeoh

THE WEB THAT WAS - RESAW 2019 | AMSTERDAM | JUNE 22, 2019

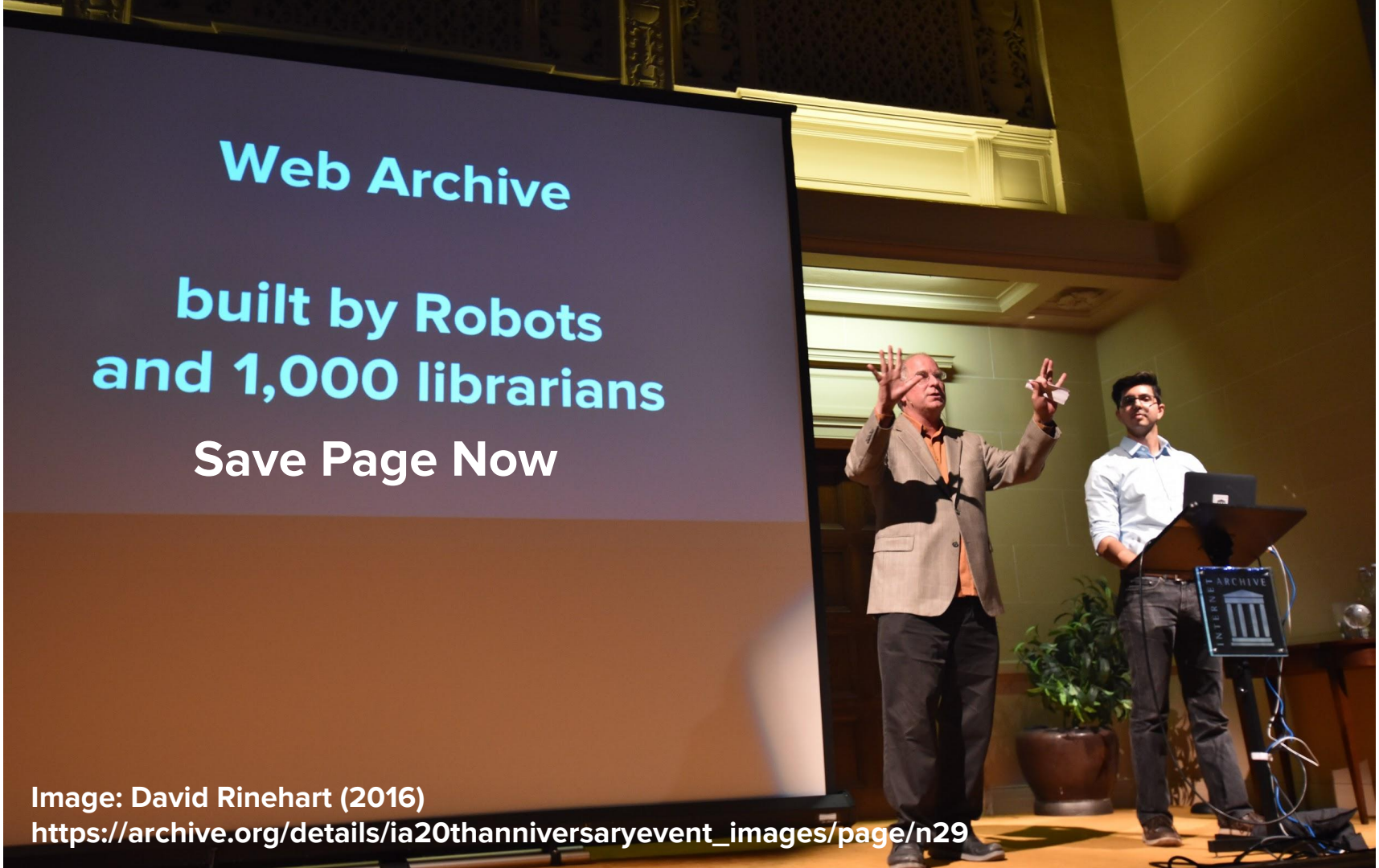


Image: David Rinehart (2016)
https://archive.org/details/ia20thanniversaryevent_images/page/n29

Brewster Kahle
@brewster_kahle

Following

651,621,510,000 web URL's now in the Wayback Machine by @internetarchive . Billions and Billions of web pages! users hitting "save page now" at 100 per second: web.archive.org

7:56 PM - 9 May 2018

68 Retweets **159** Likes

3 68 159

If You See Something, Save Something – 6 Ways to Save Pages In the Wayback Machine

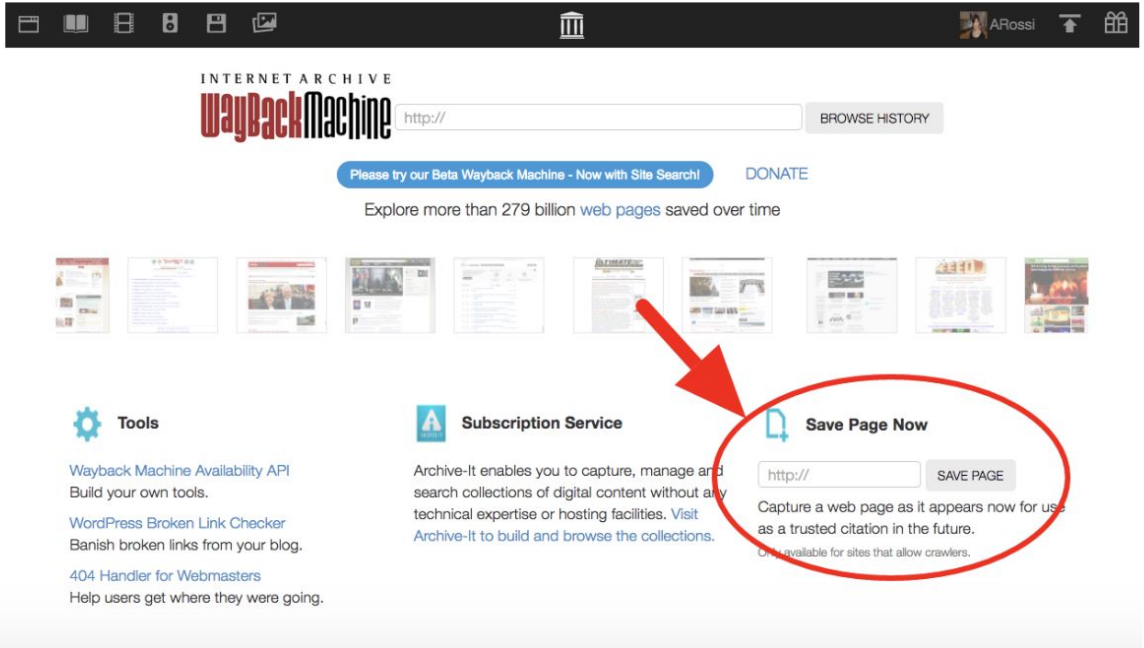
Posted on January 25, 2017 by Alexis Rossi

In recent days many people have shown interest in making sure the Wayback Machine has copies of the web pages they care about most. These saved pages can be cited, shared, linked to – and they will continue to exist even after the original page changes or is removed from the web.

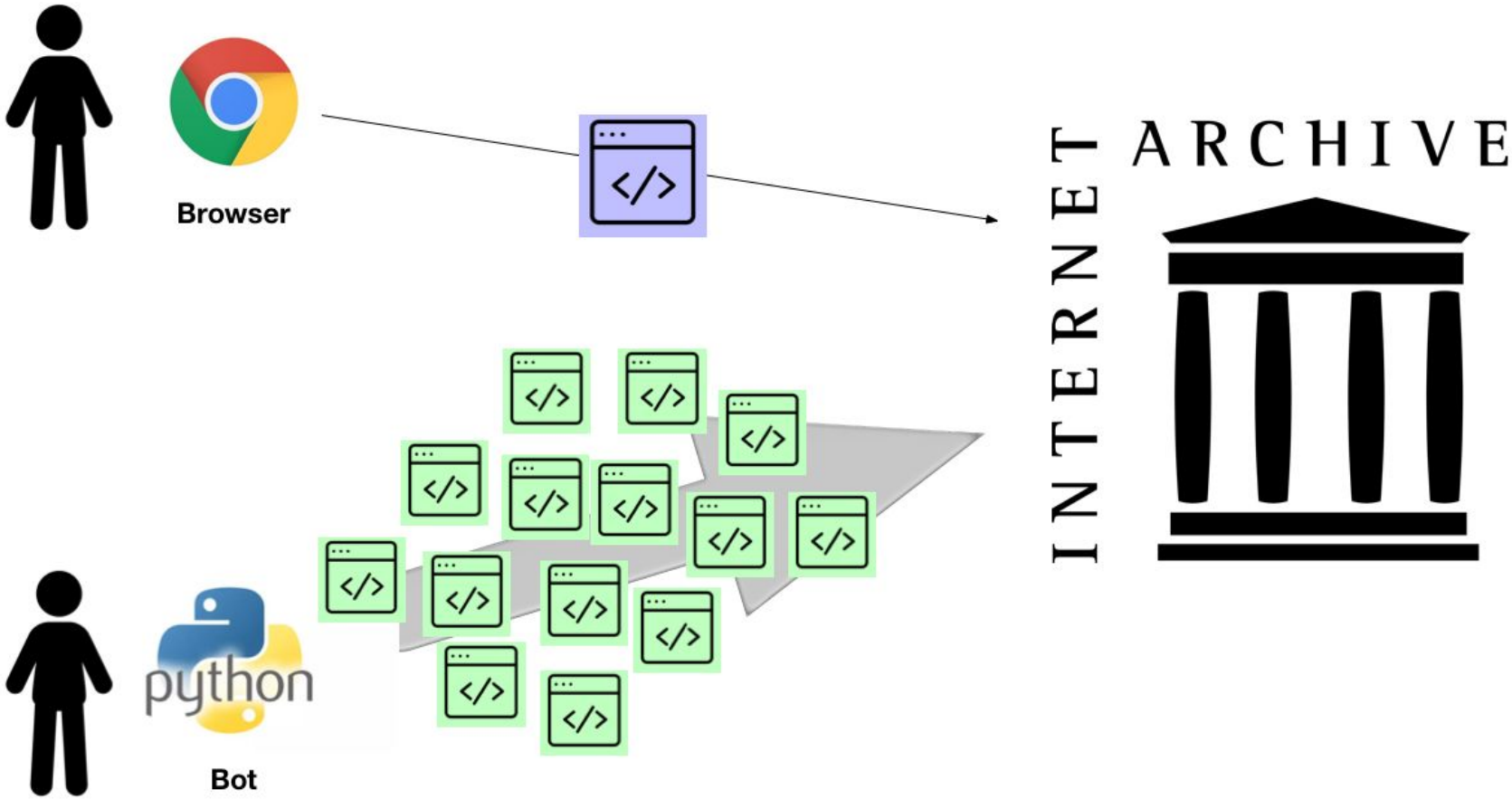
There are several ways to save pages and whole sites so that the
Machine. Here are 6 of them.

1. Save Page Now

Put a URL into the form, press the button, and we save the pag
permanent URL for your page.



https://blog.archive.org/2017/01/25/see-something-save-something/


Browser
python
Bot
INTERNET ARCHIVE

SPN has changed over time

SPN v1

- Heritrix
- *No difference between web & API submissions*

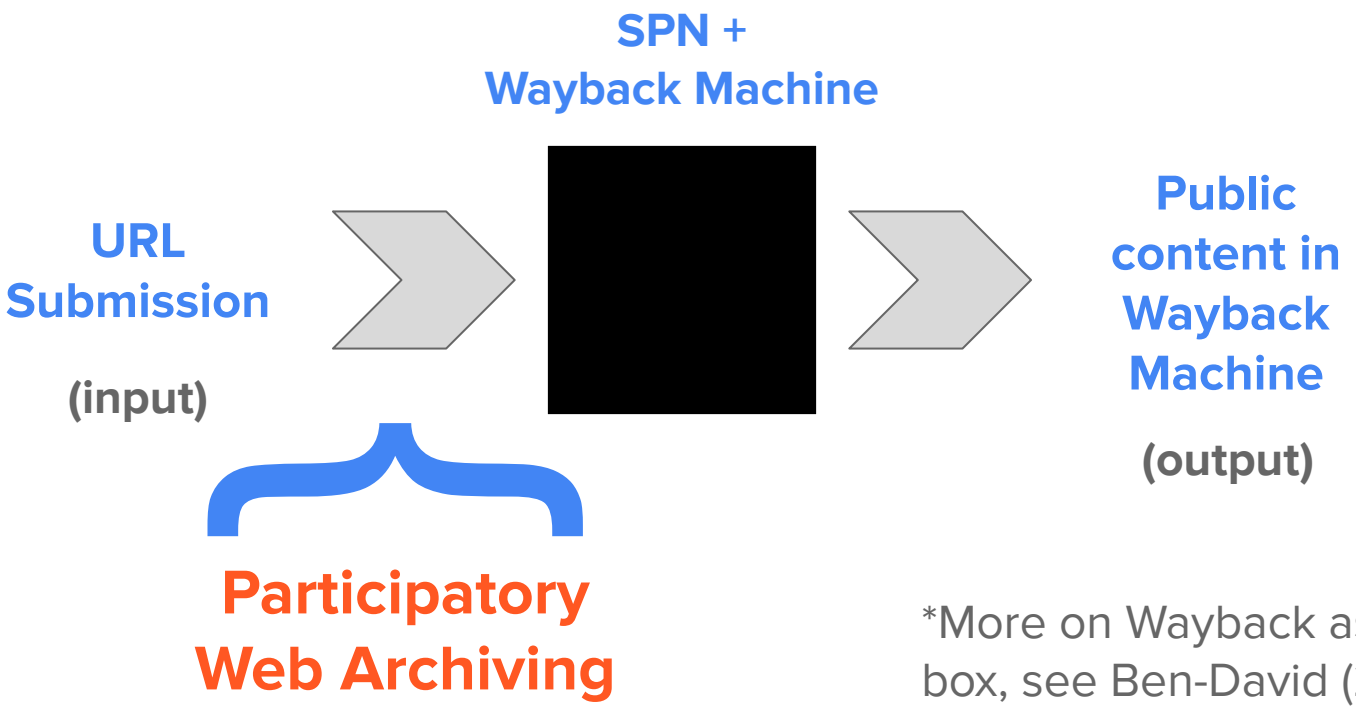
SPN v2

- Heritrix for API submissions
- Browser-based archiving for web submissions

SPN v3

- Server-based headless browser for API submissions
- Browser-based archiving for web submissions

SPN as black box



*More on Wayback as black box, see Ben-David (2018)

Motivating Research Questions

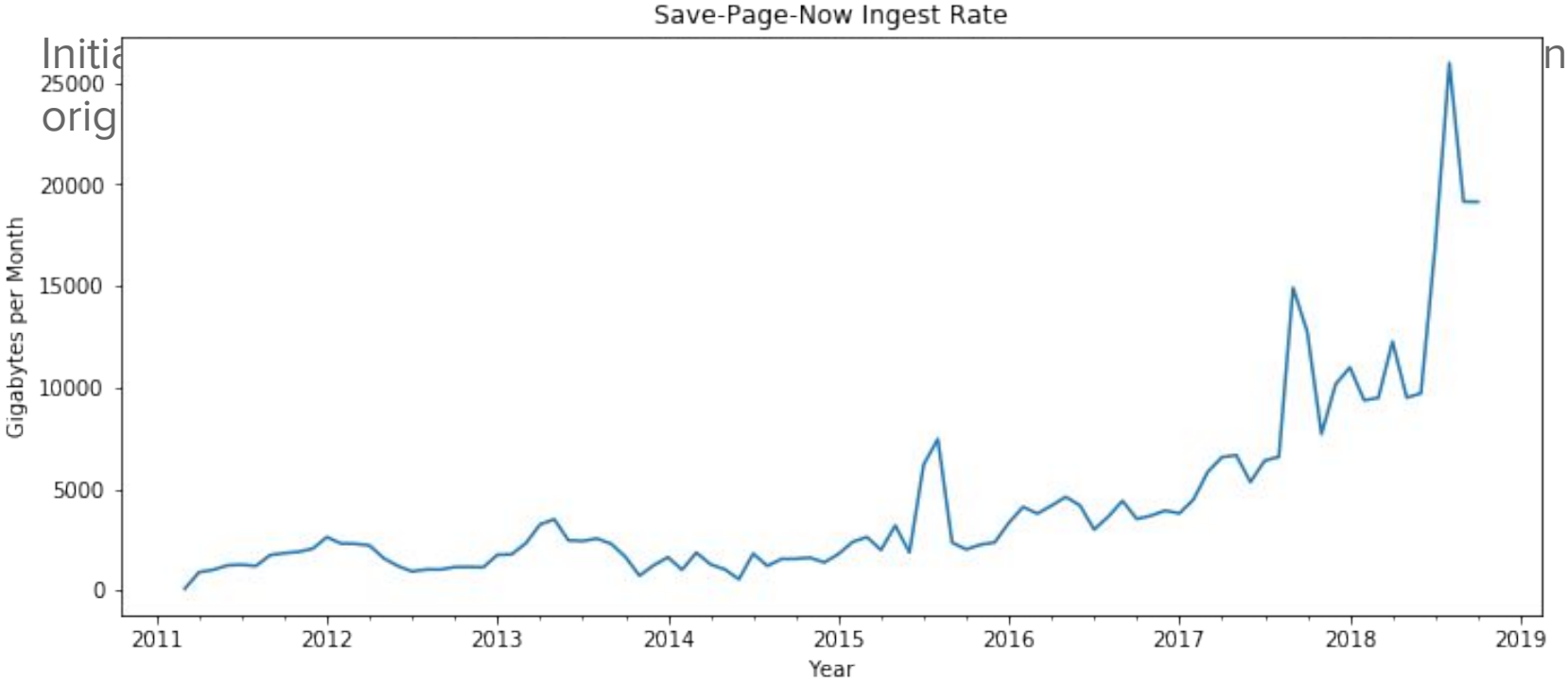
Aim: *to understand SPN as form of participatory WA infrastructure*

- **RQ1**: What is saved via SPN and how has the ‘collection’ changed over time?
- **RQ2**: To what extent are SPN resources available on the live Web and in other web archives?
- **RQ3**: In what ways is automation a factor in archival production and what purposes does it serve?

Methodology

Methodological Frame (High Level)

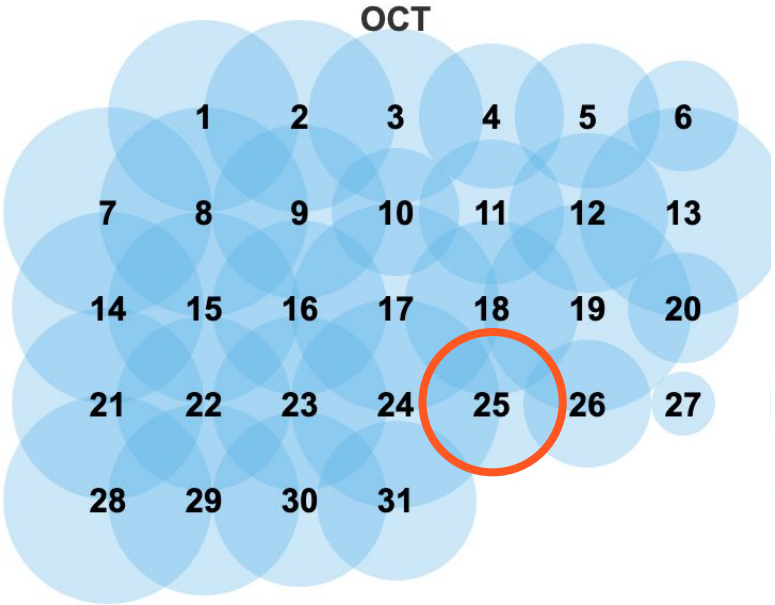
- Initia nd orig



Sample:

- One day per year
- 5 years of data
- Oct 25, 2013 - 2018

SPN first made public on homepage on October 25, 2013



WARCs downloaded via API

python

+

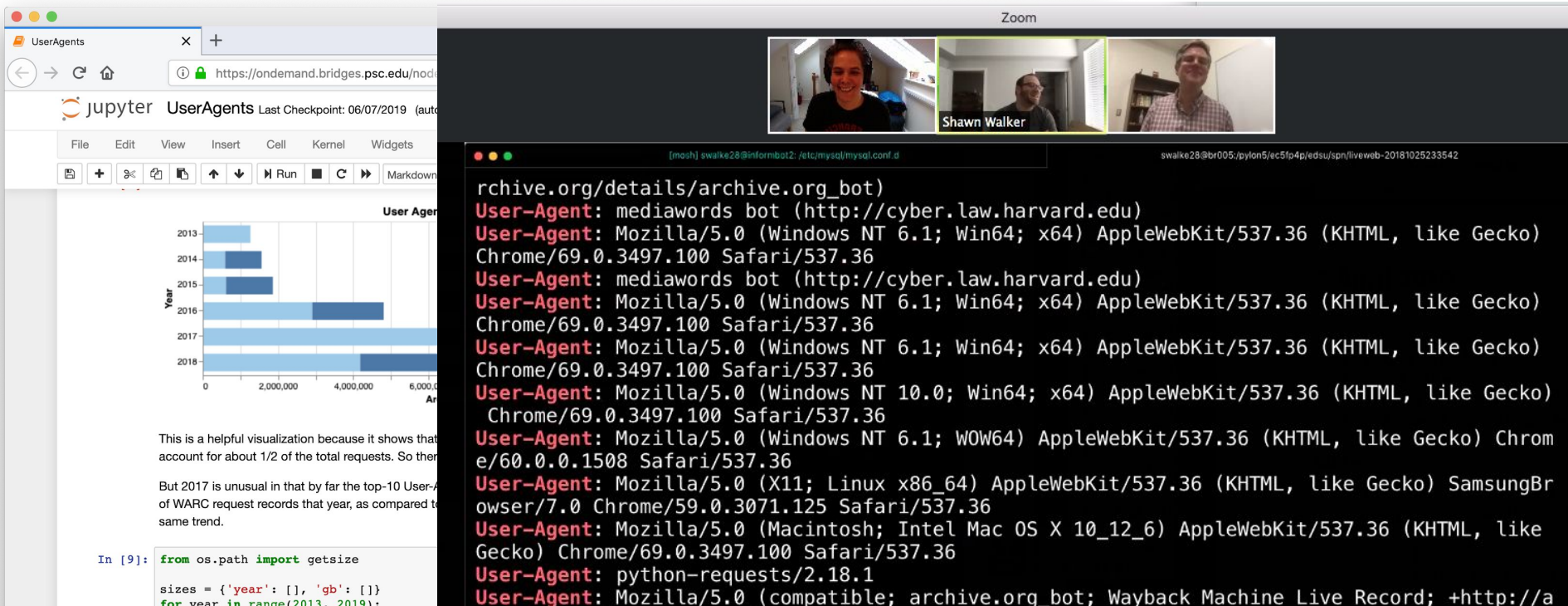
warcio

+

AUT Toolkit

Methodological Frame (High Level)

- Importance of collaborative working



RQ1: SPN Collection Development

zoominfo.

Hey guys!

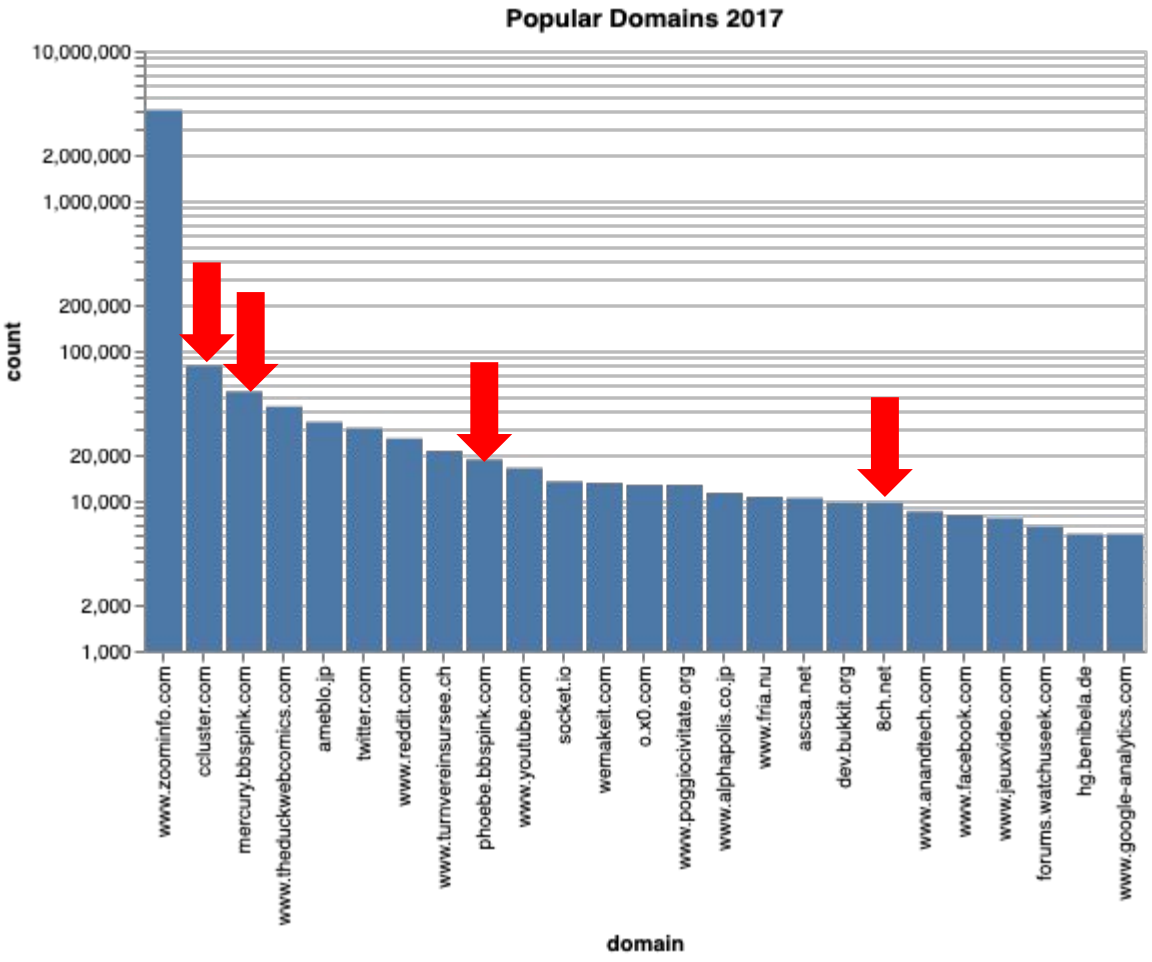
Check out my imageboard

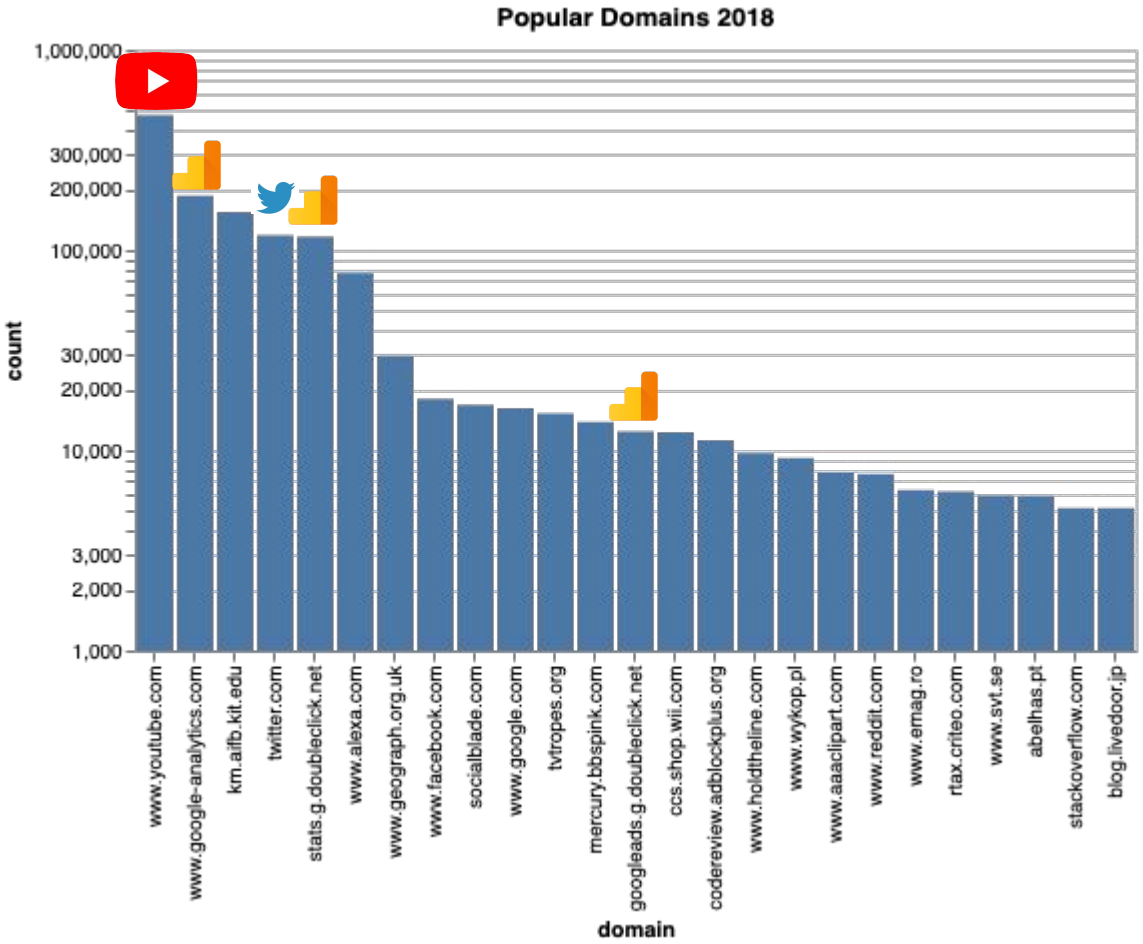
It's called 600chan

It has a /b/ board!!!!

It's also better than 4chan and 8chan combined

Plz visit my website, no one goes there...

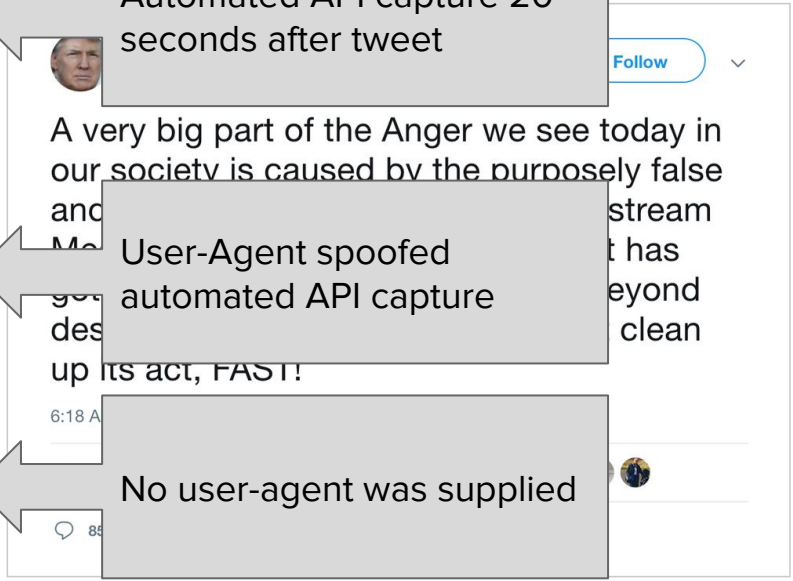



Popular Domains 2018
count
1,000,000
300,000
200,000
100,000
30,000
20,000
10,000
3,000
2,000
1,000
www.youtube.com
www.google-analytics.com
km.aifb.kit.edu
twitter.com
stats.g.doubleclick.net
www.alexa.com
www.geograph.org.uk
www.facebook.com
socialblade.com
www.google.com
tvtropes.org
mercury.bbspink.com
googleads.g.doubleclick.net
ccs.shop.wii.com
codereview.adblockplus.org
www.holdtheline.com
www.wykop.pl
www.aaaclipart.com
www.reddit.com
www.emag.ro
rtax.criteo.com
www.svt.se
abelhas.pt
stackoverflow.com
blog.livedoor.jp
domain

User-agents and Automation

	date	user_agent
0	2018-10-25T11:18:39Z	python-requests/2.13.0
1	2018-10-25T11:37:07Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
2	2018-10-25T12:00:22Z	python-requests/2.18.1
3	2018-10-25T12:01:45Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
4	2018-10-25T12:11:19Z	python-requests/2.18.1
5	2018-10-25T12:21:36Z	python-requests/2.18.1
6	2018-10-25T12:31:56Z	python-requests/2.18.1
7	2018-10-25T12:43:07Z	python-requests/2.18.1
8	2018-10-25T13:19:18Z	Chrome 41.0.2227.0
9	2018-10-25T13:54:52Z	Firefox 40.1
10	2018-10-25T13:58:42Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
11	2018-10-25T13:58:43Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
12	2018-10-25T14:32:02Z	Safari 5.1.7
13	2018-10-25T15:13:25Z	Chrome 41.0.2227.0
14	2018-10-25T15:37:12Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
15	2018-10-25T15:37:12Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
16	2018-10-25T15:53:29Z	Firefox 33.0
17	2018-10-25T16:32:46Z	Chrome 41.0.2228.0
18	2018-10-25T17:08:11Z	Chrome 41.0.2228.0
19	2018-10-25T17:44:50Z	Firefox 36.0
20	2018-10-25T18:21:49Z	Chrome 41.0.2228.0
21	2018-10-25T18:58:06Z	Safari 6.0
22	2018-10-25T19:34:05Z	Chrome 41.0.2227.1

Automated API capture 20 seconds after tweet



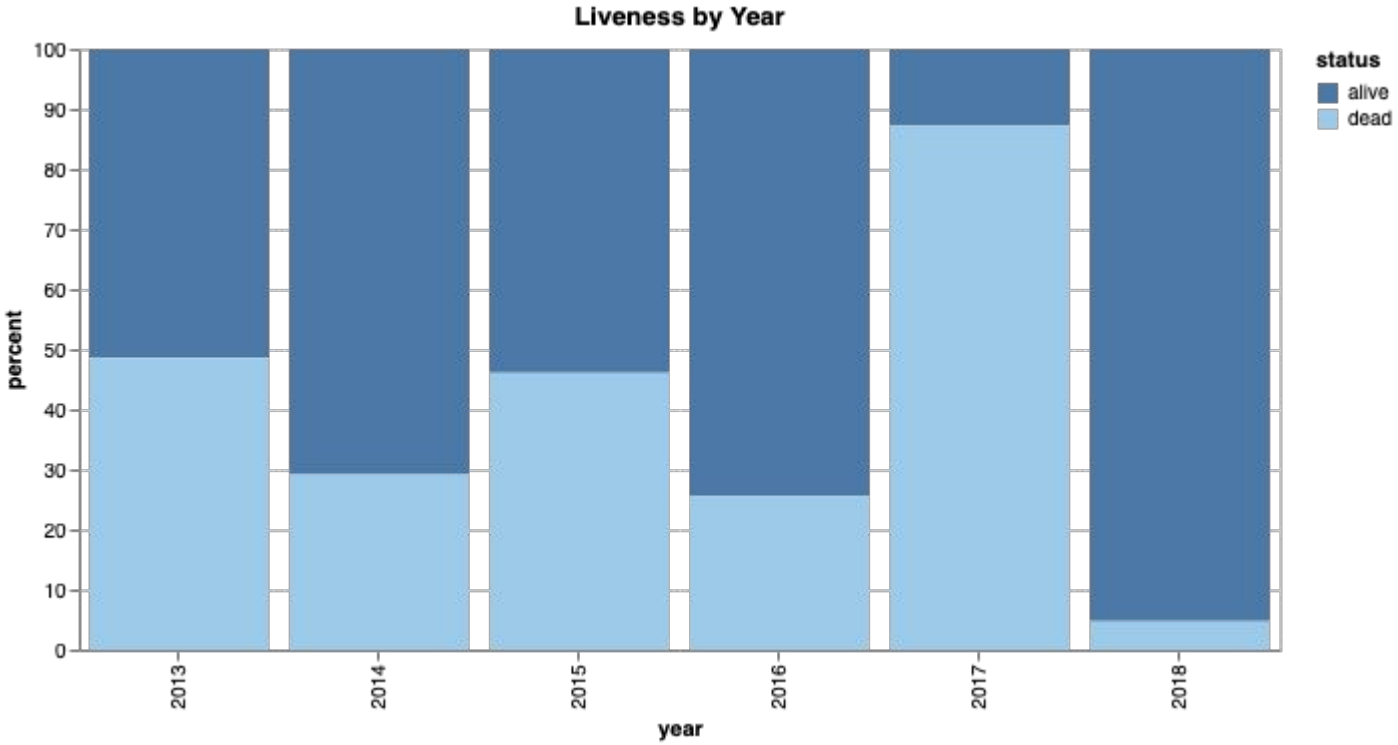
User-Agent spoofed automated API capture

No user-agent was supplied

All of these are automated captures

RQ2: Measuring Live(li)ness

- Sample of the URLs in dataset
- 2017 - ZoomInfo URLs die



RQ3: Assessing the Role of Automation

Assessing the Role of Automation

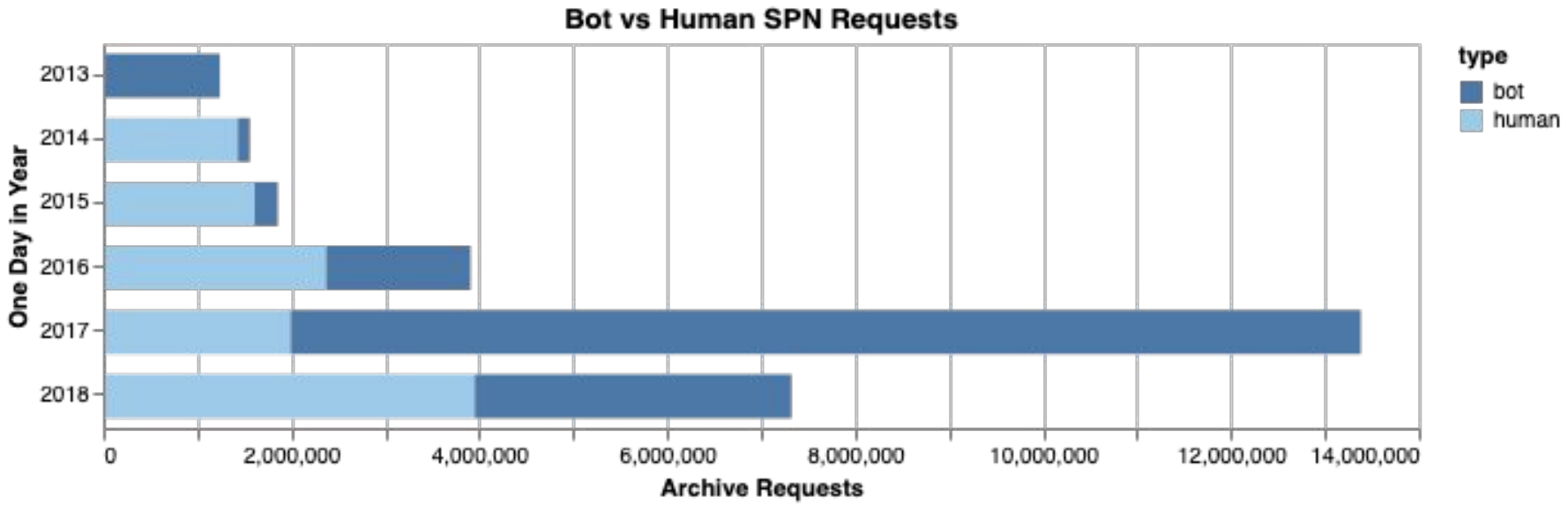
1) Hypothesised two types of SPN interactions -

- Software as **intermediary** (home page tool, browser extension) - SPN *mediates* intent of user and transports to archive
- Software as **actant** (API, cron) - SPN *transforms* selection and transports to archive

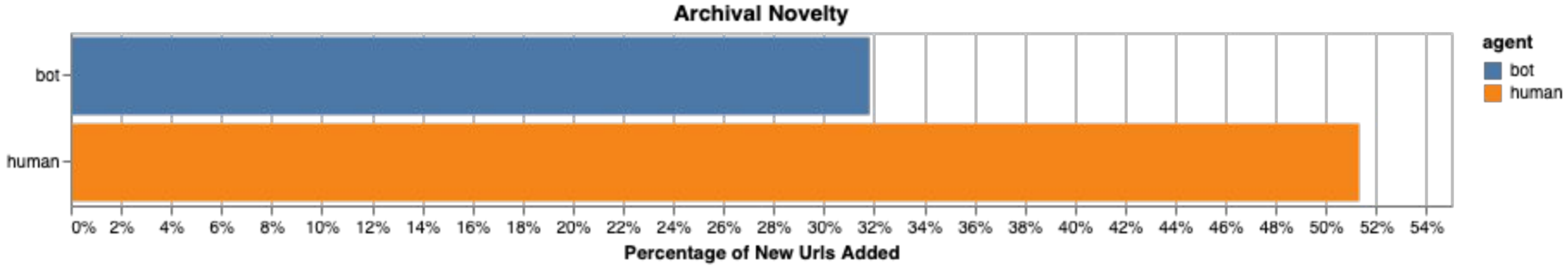
2) User-Agents as proxy for detecting automation

- User-Agents and grouping User-Agent Families
- Percentage of traffic from different types of Users

Bots vs Browsers?



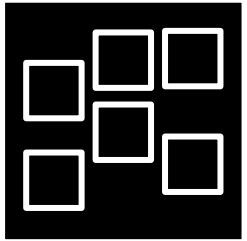
Novelty


Archival Novelty
bot
human
0% 2% 4% 6% 8% 10% 12% 14% 16% 18% 20% 22% 24% 26% 28% 30% 32% 34% 36% 38% 40% 42% 44% 46% 48% 50% 52% 54%
Percentage of New Urls Added
agent
bot
human

Limitations

- Sample strategy - it's **one day a year.**
- Detecting automation
- Conceptualizing web archival attributes (e.g. 'novelty')
- URLs - delineating targets ➔ inferring intentionality

Reflexivity as Strategy



- Complex layering of data and findings (boxes within boxes), ways that proxies/data abstraction often creates more questions + uncertainty
- Problematizing *situatedness* (Haraway 1988) of so-called big data (+ extra complexities of SPN) - recognizing that data views are always partial, '*cooking data with care*' (Bowker 2005; Geiger and Halfaker 2018)
- Value of triangulation, epistemological flexibility - WA research as *boundary work* (Star and Griesemer 2015; Gieryn 1983) - requires a lot of translation

Future work...

Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges and Bridges Storage at the Pittsburgh Supercomputing Center through allocation TG-ECS180012. Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.