slides.md

The web is terrifying!

Sarah Bird

moz://a

bokeh.pydata.org

Committed to

an internet that

- includes all the peoples of the earth
- promotes civil discourse and individual expression
- elevates critical thinking and shared knowledge
- catalyzes working together for the common good

Mozilla Manifesto & 2018 Addendum

Tracking technologies

Cookies

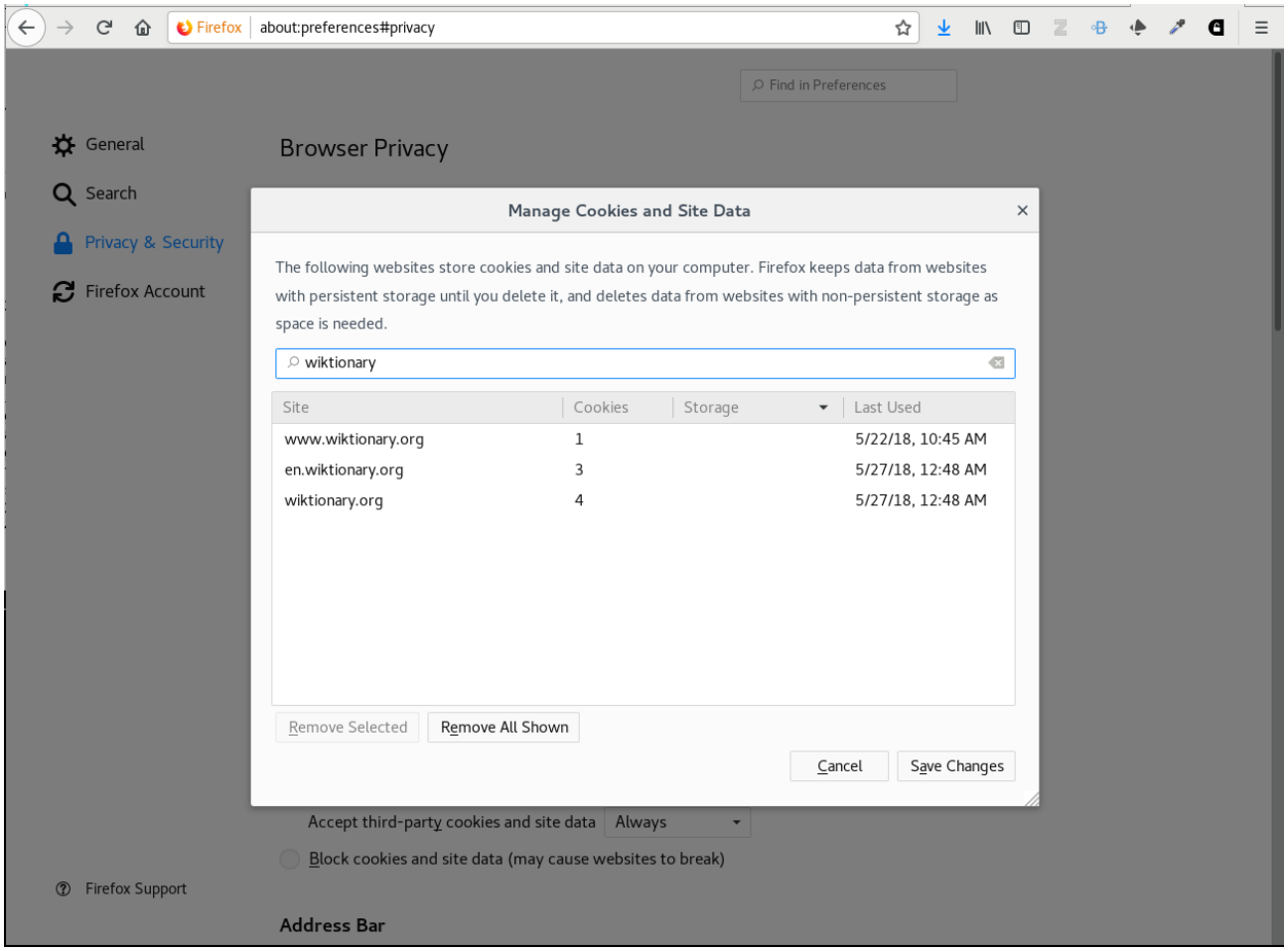
> An HTTP cookie is a small piece of data that a server sends to the user's web browser. The browser may store it and send it back with the next request to the same server. developer.mozilla.org/en-US/docs/Web/HTTP/Cookies

---

> tmpPersistentuserId

> 43038346ae23c38666a69e44f980f244

> haaretz.com



In [11]:
```
df[df.baseDomain.str.contains('wiktionary')][['host', 'name', 'value']]
```

Out[11]:

	host	name	value
2976	www.wiktionary.org	WMF-Last-Access	22-May-2018
3801	en.wiktionary.org	WMF-Last-Access	26-May-2018
3802	.wiktionary.org	WMF-Last-Access-Global	26-May-2018
3818	en.wiktionary.org	enwiktionaryUserID	
3819	en.wiktionary.org	enwiktionaryUserName	Birdsarah
3820	.wiktionary.org	forceHTTPS	true
3821	.wiktionary.org	centralauth_User	Birdsarah
3822	.wiktionary.org	centralauth_Token	

Let's take a look

rubiconproject.com - 88 cookies

define rubicon

> *noun* A point of no return.

> *example* Once you've crossed the Rubicon there's no going back.

3 months

> 5700 cookies

disconnectme/disconnect-tracking-protection

Categorization from disconnect.me

category	count
Advertising	914
Content	292
Analytics	112
Social	55
Disconnect	38
Uncategorized*	4303

category count

* Not identified by disconnect as belonging to a particular tracking category. We'll come back to this later

But you can delete your cookies.

Zombie cookie & Cookie sync

Cookie Syncing

> "the process by which two different trackers link the IDs they've given to the same user" freedom-to-tinker.com/2014/08/07/the-hidden-perils-of-cookie-syncing/

> "This guide explains how the Cookie Matching Service enables you to make more effective bidding choices." developers.google.com/ad-exchange/rtb/cookie-guide

---

Any evidence of syncing?

Interlude - my learning journey:

- Python, JS, stuff - I learned as a software engineer
- Pandas - https://github.com/brandon-rhodes/pycon-pandas-tutorial
- Bokeh

Ingredients:

- Publications
- People
- Projects

Zombie cookie

> a cookie that is recreated after deletion

> by storing the cookie data in several types of storage mechanisms that are available on the browser. samy.pl/evercookie/

Any evidency of zombies?

Part 1 Part 2

> "A cookie is a small file placed onto your device that enables LinkedIn features and functionality." LinkedIn

> *and enables companies to compile your browsing history and online behavior*

Do you care who knows your browsing history?

What about browsing history + location + smartfridge + drone feeds?

What about if that data is is stolen?

What if you paid 20% more for a service than your friend?

> Your browsing history can be, and is, connected to your real-world identity

---

twitter, reddit, github, instagram, linkedin, ...

---

acxiom

The World's Most Advanced Omnichannel Identity Graph

We recognize consumers across offline and digital touchpoints at scale in a privacy-compliant way, enabling you to gain the most comprehensive, omnichannel understanding of your customers.

---

cookies -> browsing history

Okay, I'll disable cookies.....

---

browsing history -> identity

The language we love to hate

JavaScript

Systems Research Group

OverScripted!

---

1 million locations visited

131m javascript calls recorded

Systems Research Group

+

UCOSP students

---

https://hacks.mozilla.org/2018/06/overscripted-digging-into-javascript-execution-at-scale

---

This is where I came into the data

Downloading, processing, cleaning the data

> **dask dask.pydata.org**

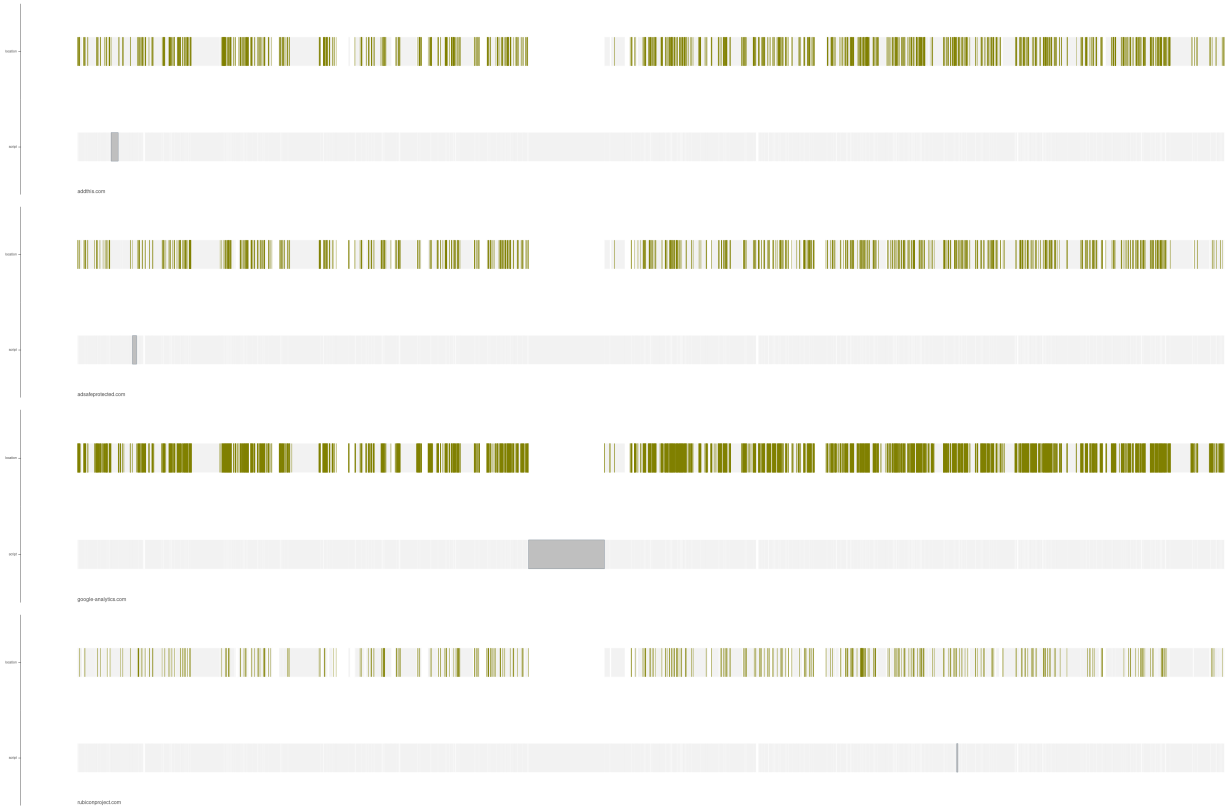
> **pyspark spark.apache.org**

```
filesRDD = sc.wholeTextFiles(BUCKET_NAME, minPartitions=3000)
```

```
>>> import dask.dataframe as dd
>>> ddf = dd.read_parquet('../cache_new.parquet/')
>>> ddf.dtypes

arguments          object
call_id            object
call_stack         object
crawl_id            int64
func_name          object
in_iframe            bool
location           object
operation          object
script_col         object
script_line        object
script_loc_eval    object
script_url         object
symbol             object
time_stamp         object
value              object
dtype: object
```

First aha moment



Tracking Technologies

Canvas fingerprinting

Canvas fingerprinting

Fingerprinting

panopticlick.eff.org

Browser Characteristic	bits of identifying information	one in x browsers have this value
Limited supercookie test	0.39	1.31
Hash of canvas fingerprint	9.55	751.11
Screen Size and Color Depth	2.46	5.5
Browser Plugin Details	8.18	289.16
Time Zone	4.25	18.99
DNT Header Enabled?	1.26	2.39
HTTP_ACCEPT Headers	2.0	4.01
Hash of WebGL fingerprint	9.83	909.19
Language	0.92	1.89
System Fonts	14.62	25121.65
Platform	3.25	9.5
User Agent	13.08	8639.72
Touch Support	0.59	1.51
Are Cookies Enabled?	0.22	1.16

Categorization from disconnect.me

category	count
Advertising	919
Content	263
Analytics	105
Social	51
Disconnect	37
Uncategorized	4114

Fingerprinting the fingerprinters

A work in progress to see if we can use machine learning to automate the detection of tracking technlogies.

Any hope?

	Technology

	Policy
	Community

Technology

- Firefox
- Private search
- Privacy add ons
- VPN
- Tor (network and browser)

Policy

GDPR

European Union's General Data Protection Regulation

Community

- educate yourself and others
- activism
- get involved in local efforts
- stand up for your rights
- complain on twitter

get digging

github.com/mozilla/Overscripted-Data-Analysis-Challenge

Prizes

top 3 analyses

present at MozFest 2018 in London

(airfare, hotel, admission and, if necessary, visa fees covered)

Thanks!

@birdsarah

github.com/mozilla/Overscripted-Data-Analysis-Challenge