Privacy Policies Through Time

A Dataset for Privacy Policy History Analysis and Exploration

Jack Workman
11 April 2019
W231 Section 2

Outline

- Why?
- The Privacy Policies Through Time Dataset
- Dataset Metadata
- Methodology
- Key Privacy Event Analysis
- Case Study: Google
- Next Steps
- Questions

Why?

Why?

Studying how privacy policies have evolved over time can tell us about...

1. A company's approach to privacy
2. The immediate and lasting impact of legislation
3. How to best form new legislation

Why?

Other privacy policy datasets cover only recent policies and do not connect revisions

Dataset	Policy Count	Specialty
OPP-115 Corpus (ACL 2016)	115	Annotated website policies
APP-350 Corpus (PETS 2019)	350	Annotated Android app policies
Opt-out Choice Dataset (EMNLP 2017)	102	Policies with opt-out choice labels
ACL/COLING 2014 Dataset	1010	Website policies spanning 12/2013 - 01/2014

Why?

Dataset	Policy Count	Specialty
OPP-115 Corpus (ACL 2016)	115	Annotated website policies
APP-350 Corpus (PETS 2019)	350	Annotated Android app policies
Opt-out Choice Dataset (EMNLP 2017)	102	Policies with opt-out choice labels
ACL/COLING 2014 Dataset	1010	Website policies spanning 12/2013 - 01/2014
Privacy Policies Through Time	**295**	**History and connected revisions**

The Privacy Policies Through Time Dataset

The Privacy Policies Through Time Dataset

A collection of privacy policies organized by company and revision date

Facebook

- 2005-06-08.txt
- 2006-02-27.txt
- 2006-05-22.txt

...

- 2018-04-19.txt

Apple

- 2000-10-13.txt
- 2001-05-04.txt
- 2001-06-04.txt

...

- 2018-05-22.txt

Amazon

- 2003-04-03.txt
- 2005-07-20.txt
- 2005-10-27.txt

...

- 2017-08-29.txt

And many more!

Dataset Metadata

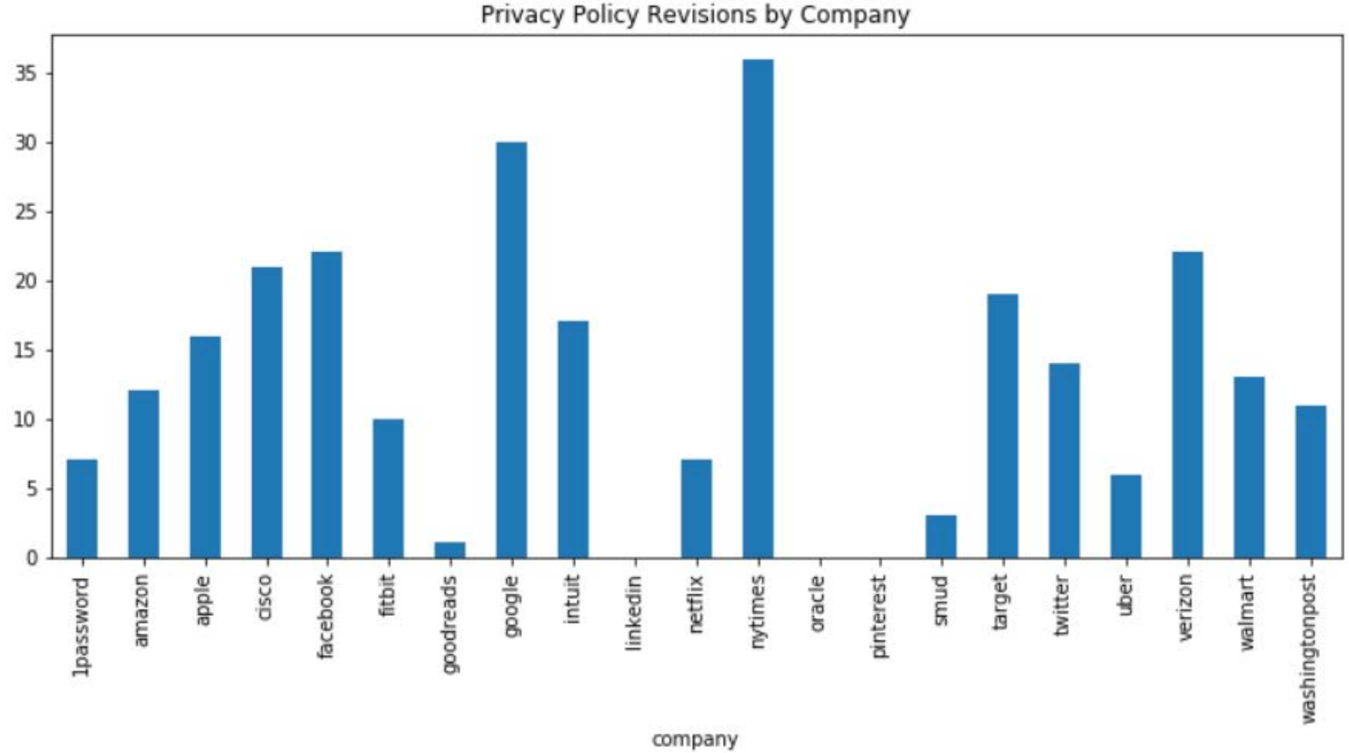
Number of Privacy Policies: 295

Number of Companies: 21

Date Range: June 9th, 1999 - January 1st, 2019

1password
Amazon
Apple
Cisco
Facebook
Fitbit
Glassdoor
Goodreads
Google
Intuit
Linkedin
Netflix
New York Times
Oracle
Pinterest
SMUD
Target
Uber
Verizon
Walmart
Wash. Post

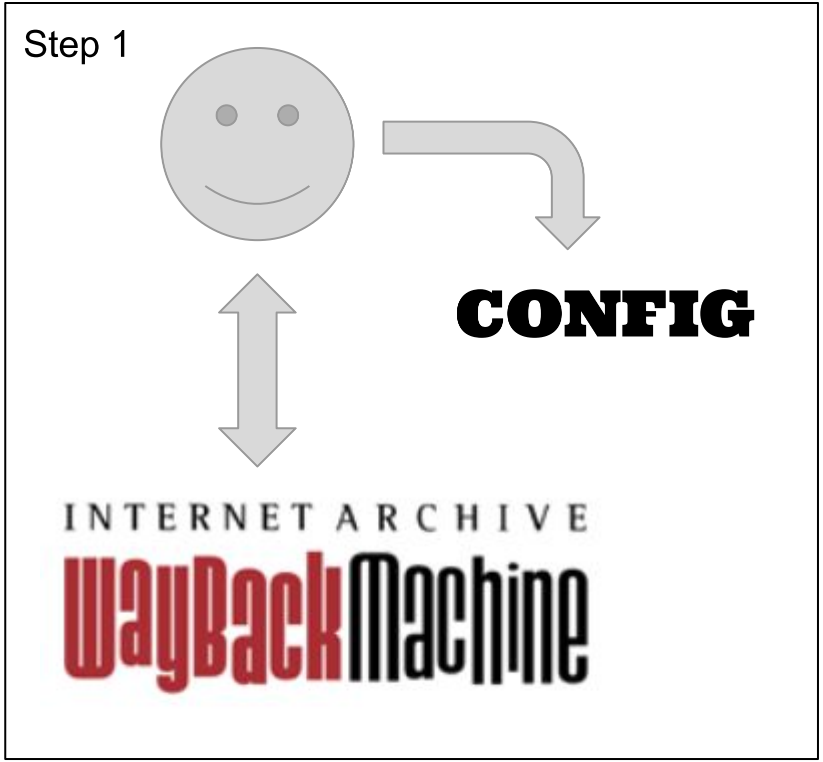
Dataset Metadata



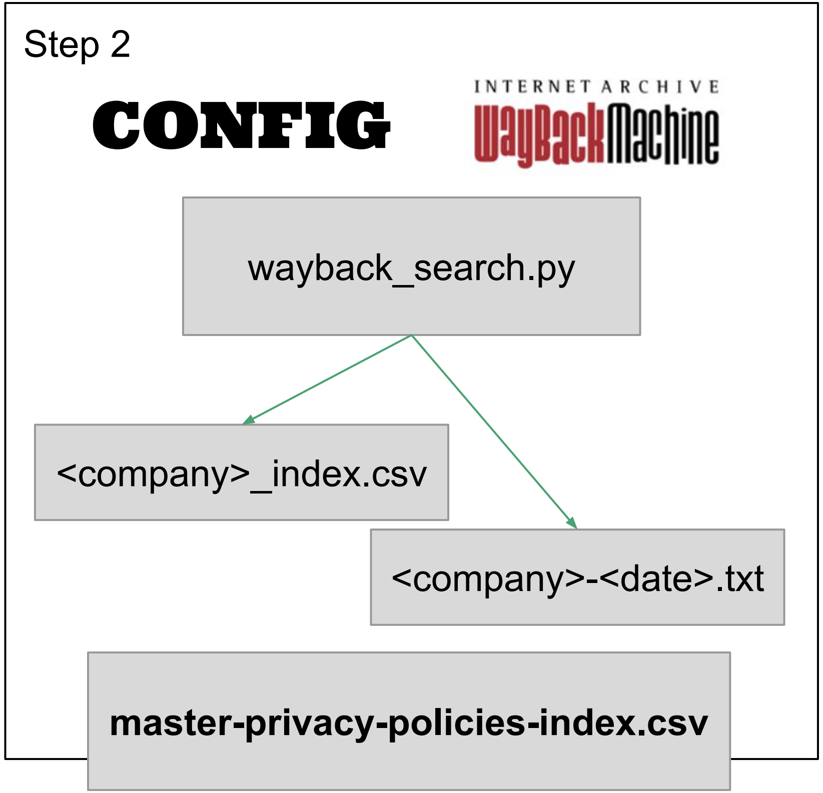
Methodology

INTERNET ARCHIVE
WayBackMachine

Methodology



Step 1



Step 2

wayback_search.py

<company>_index.csv

<company>-<date>.txt

master-privacy-policies-index.csv

Key Privacy Event Analysis

Key Privacy Event Analysis

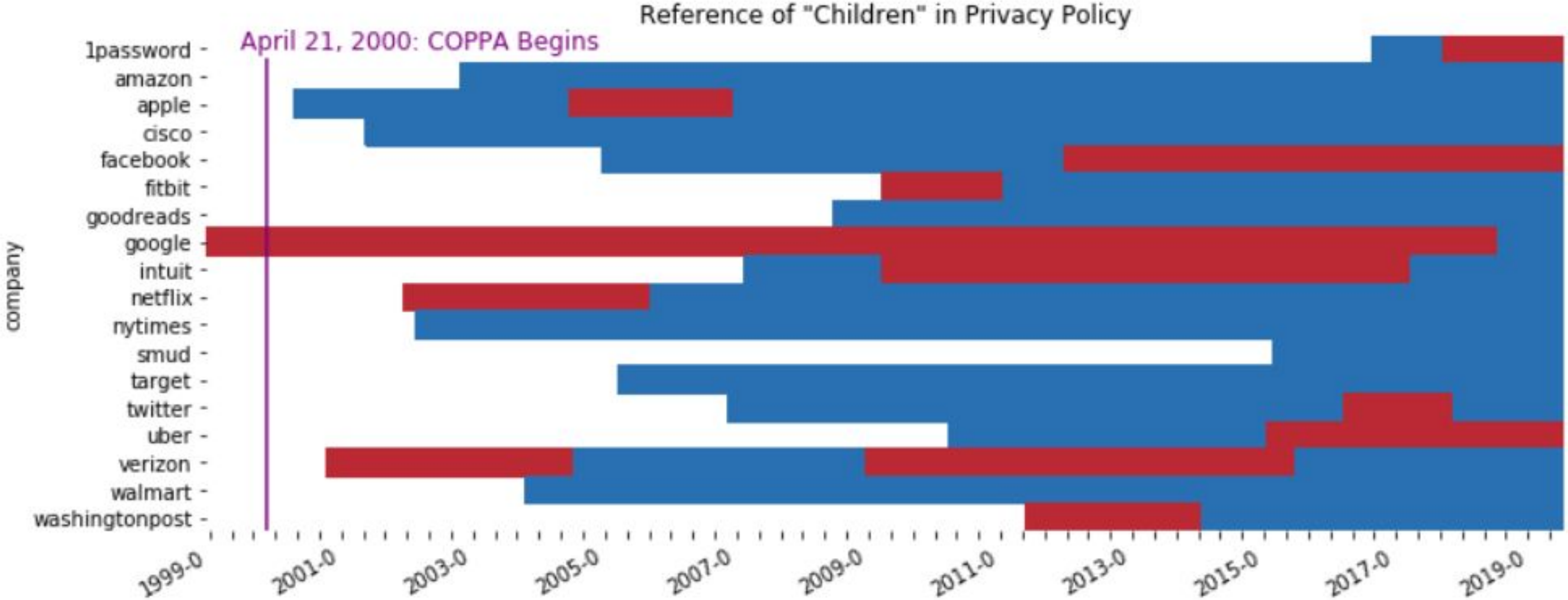
As we’ve learned in class, there have been several big pieces legislation/policy that have shaped how businesses approach privacy

- Children's Online Privacy Protection Act (COPPA)
- International Safe Harbor Privacy Principles
- EU-US Privacy Shield
- Do Not Track
- General Data Protection Regulation (GDPR)

We can explore how companies reacted to each of these with the Privacy Policies Through Time Dataset

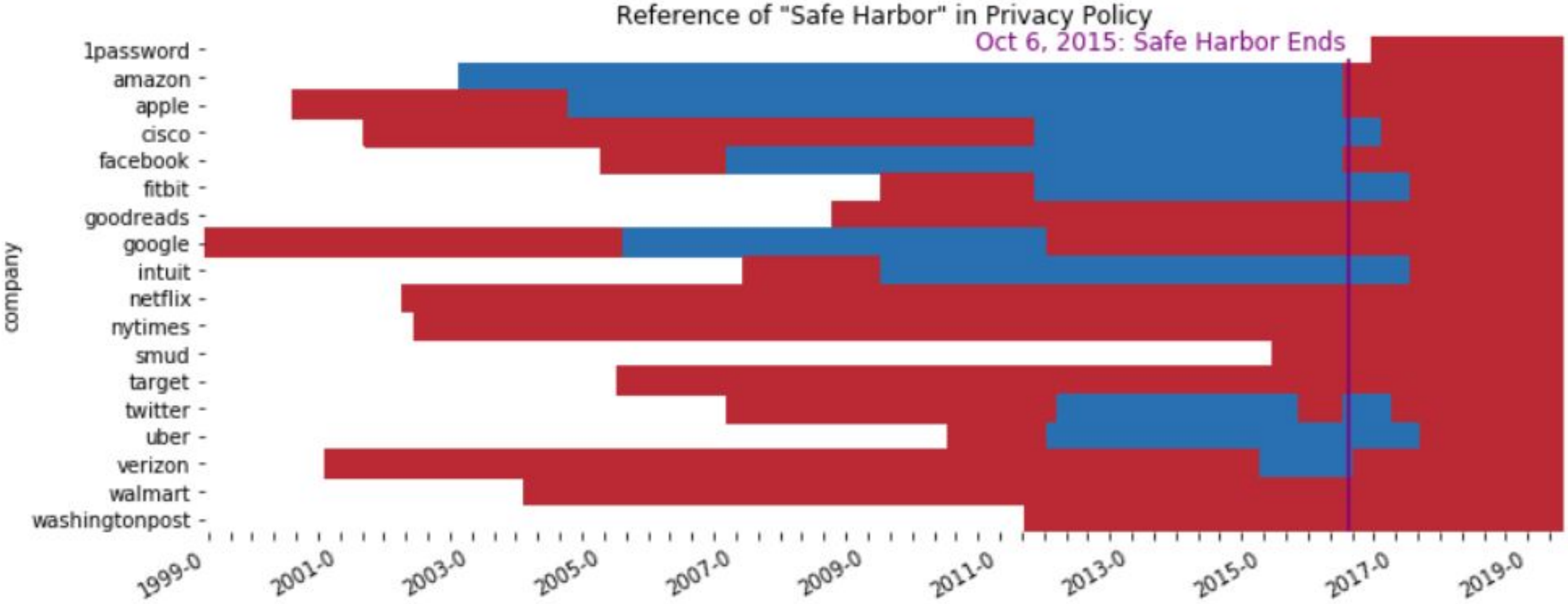
Key Privacy Event Analysis

Children's Online Privacy Protection Act (COPPA)



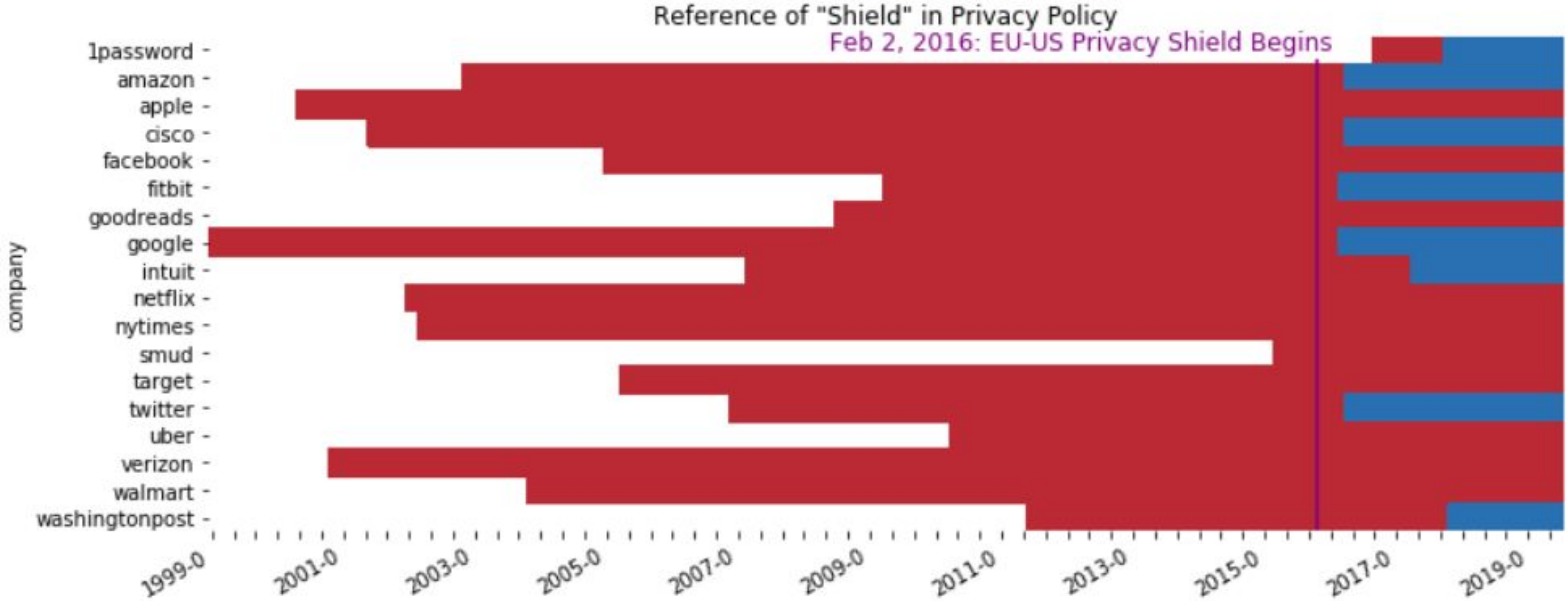
Key Privacy Event Analysis

International Safe Harbor Privacy Principles



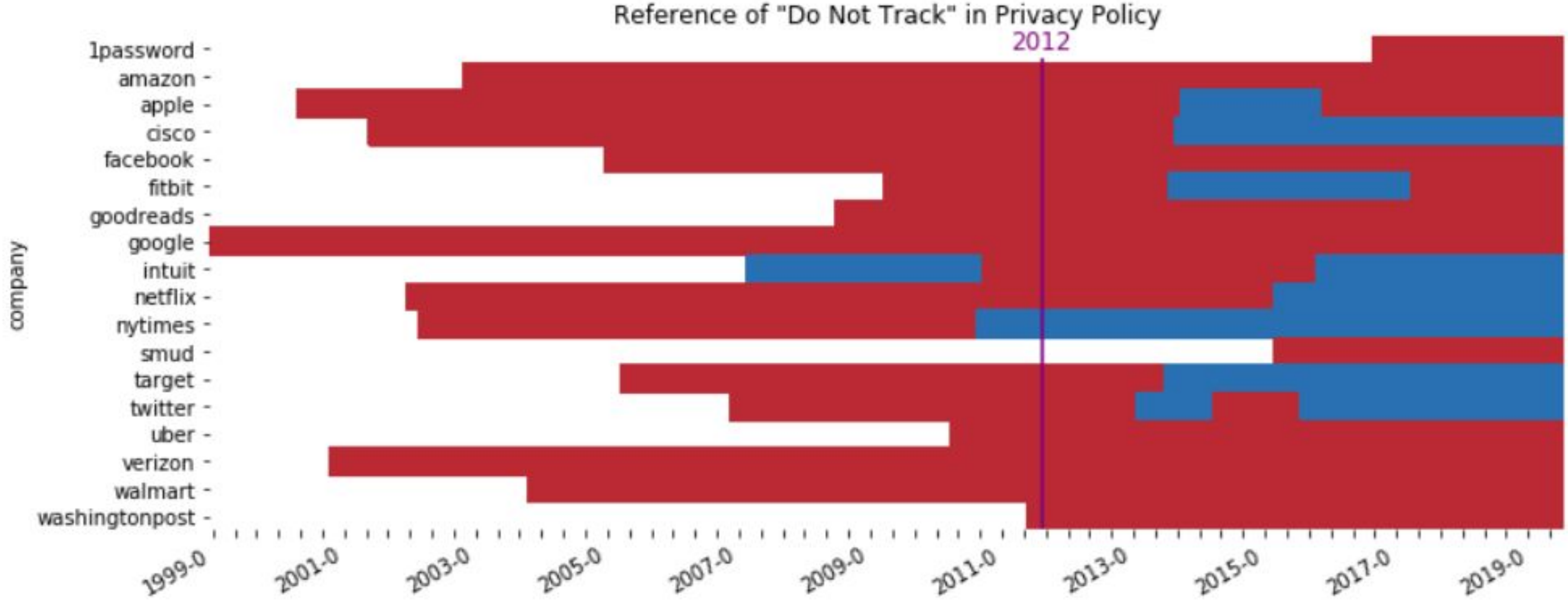
Key Privacy Event Analysis

EU-US Privacy Shield



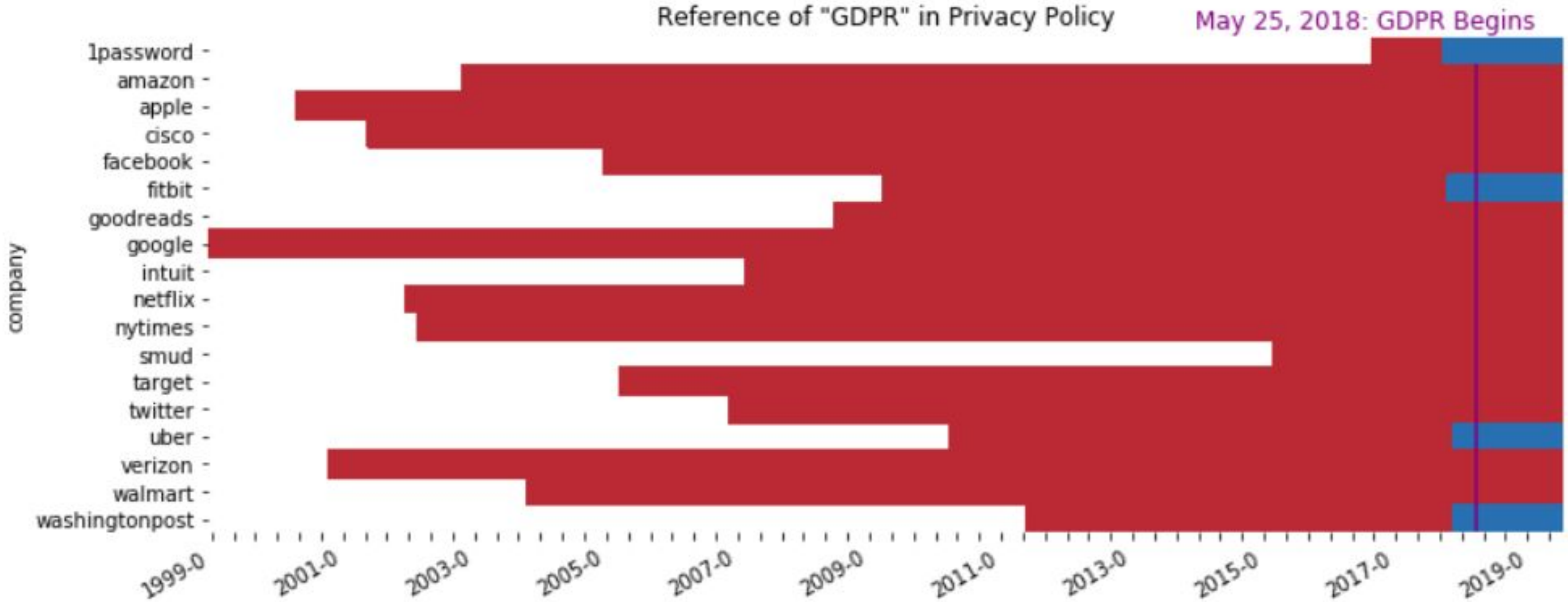
Key Privacy Event Analysis

Do Not Track



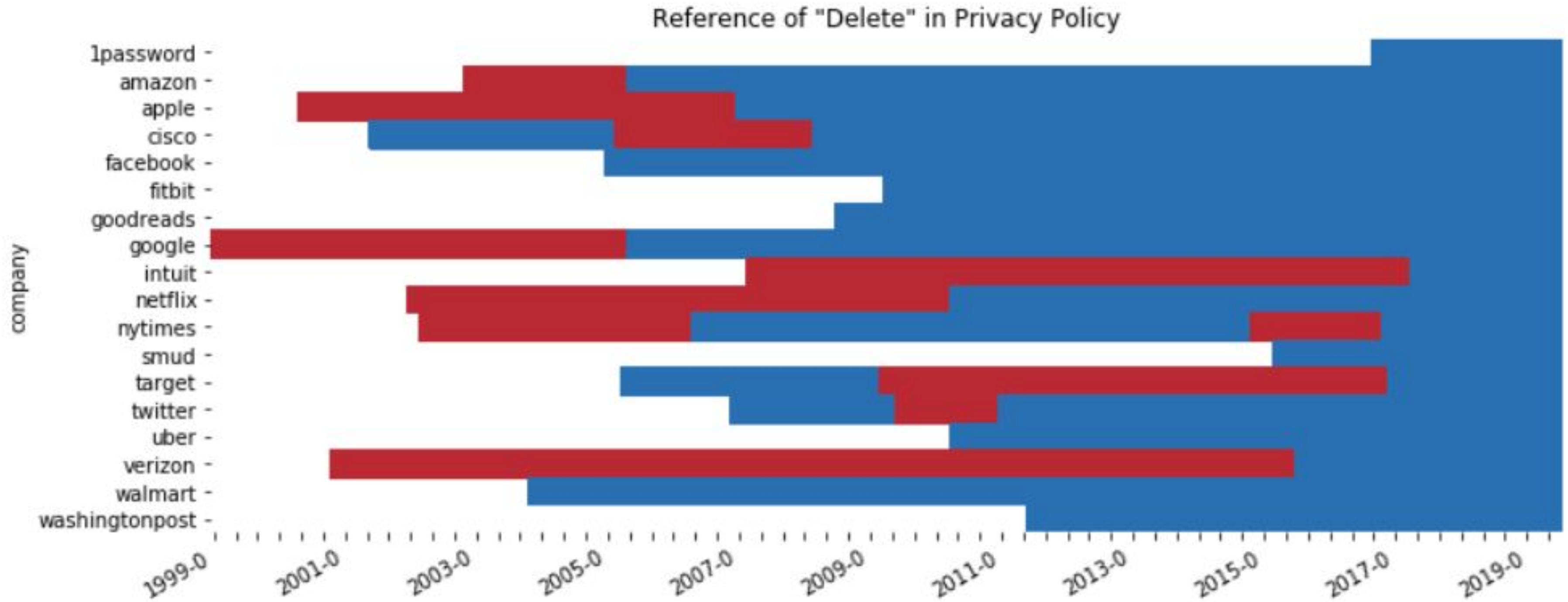
Key Privacy Event Analysis

General Data Protection Regulation (GDPR)



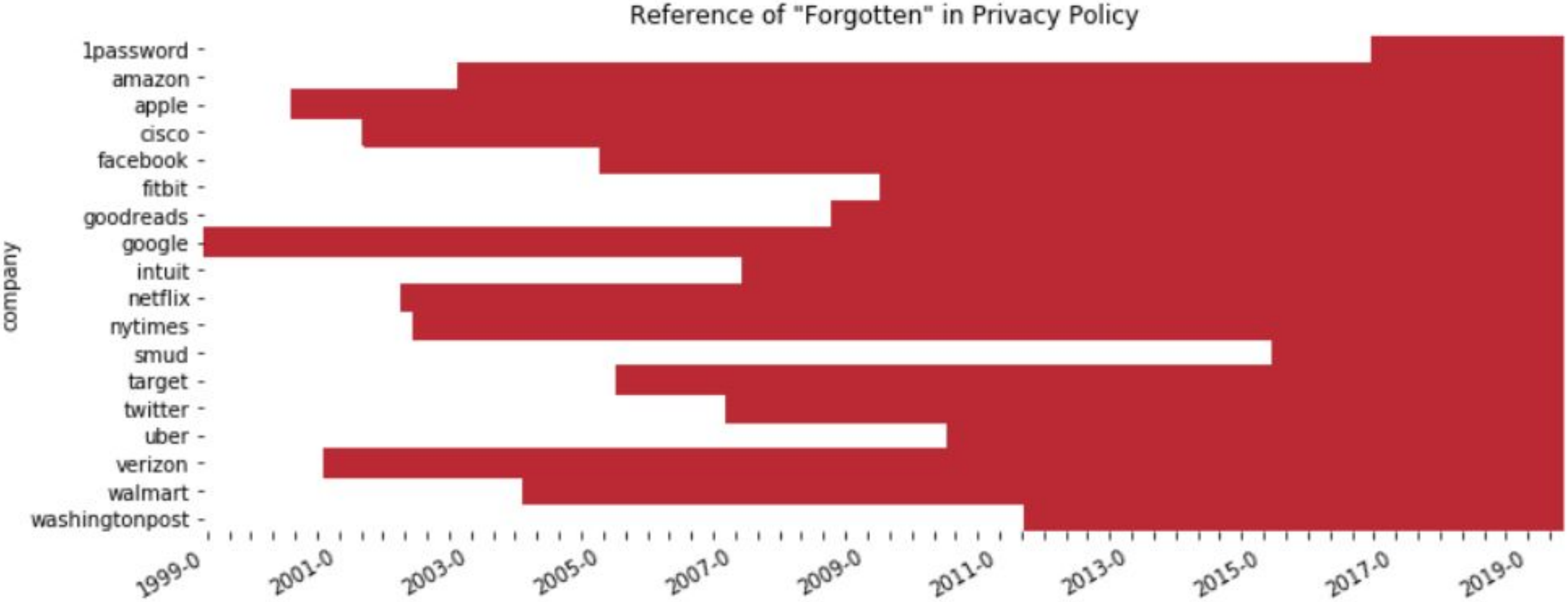
Key Privacy Event Analysis

General Data Protection Regulation (GDPR) - “Delete”?



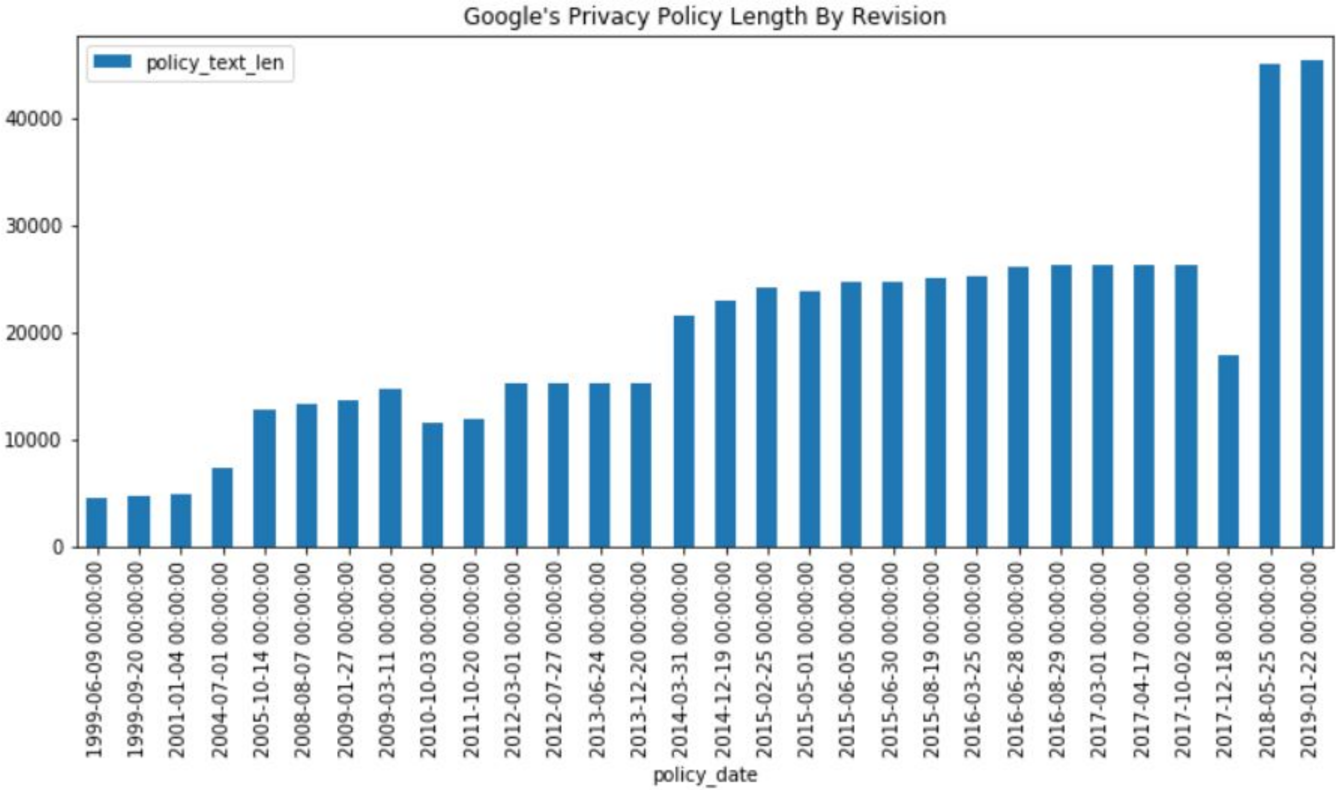
Key Privacy Event Analysis

General Data Protection Regulation (GDPR) - Right to be “Forgotten”?



Next Steps

Case Study: Google



Next Steps

Readability Metrics

1. Lexicon Count
2. Syllables Count
3. Sentence Count
4. Passive Voice Index
5. Flesch Kincaid Grade
6. Dale-Chall Readability Score

Linden, Thomas, Hamza Harkous, and Kassem Fawaz. "The Privacy Policy Landscape After the GDPR." arXiv preprint arXiv:1809.08396 (2018).

Questions

1. What would you do with this data? How would you analyze it?

2. Should I gather additional data? Any companies/industries that are missing?

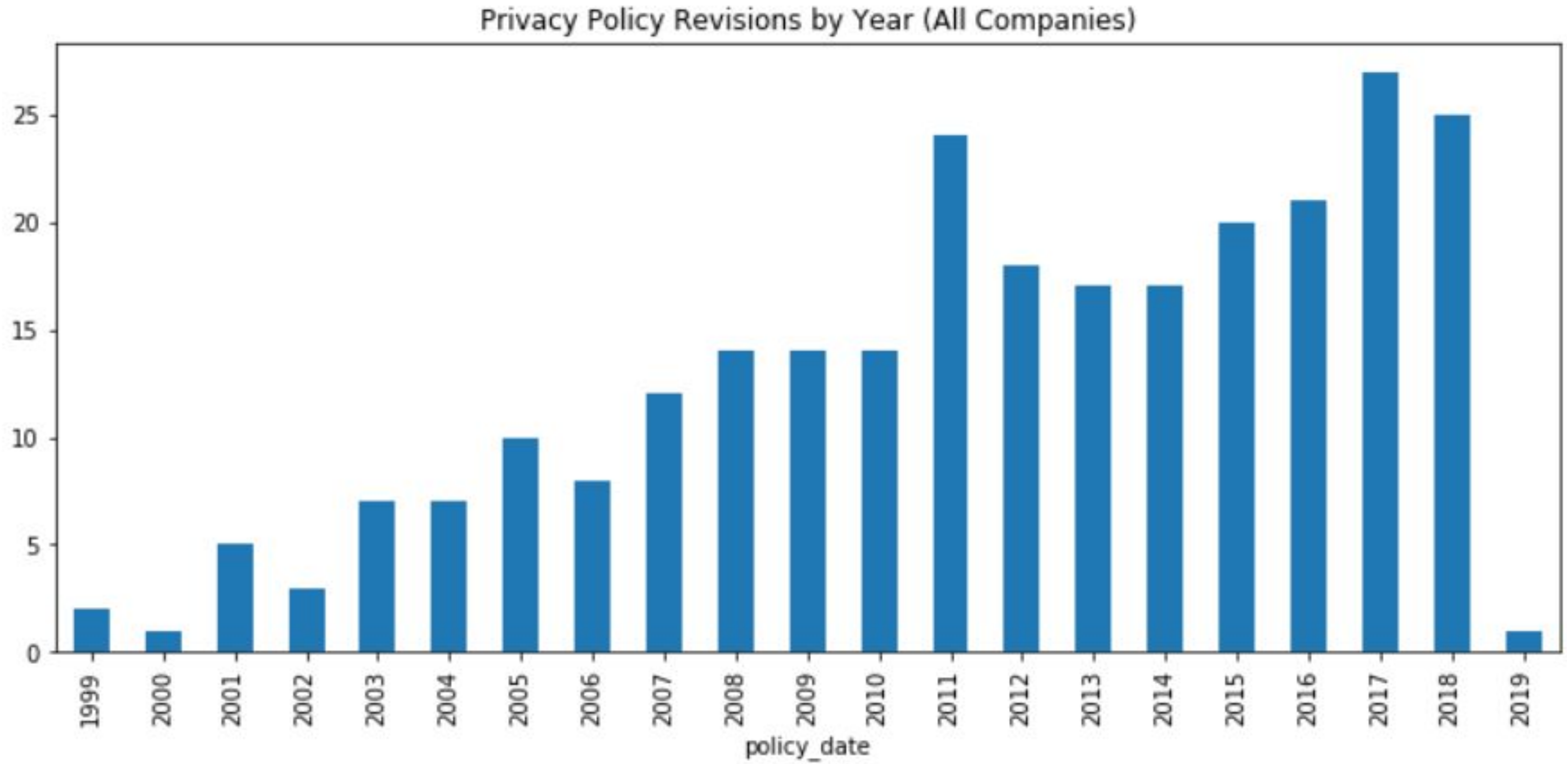
Sources

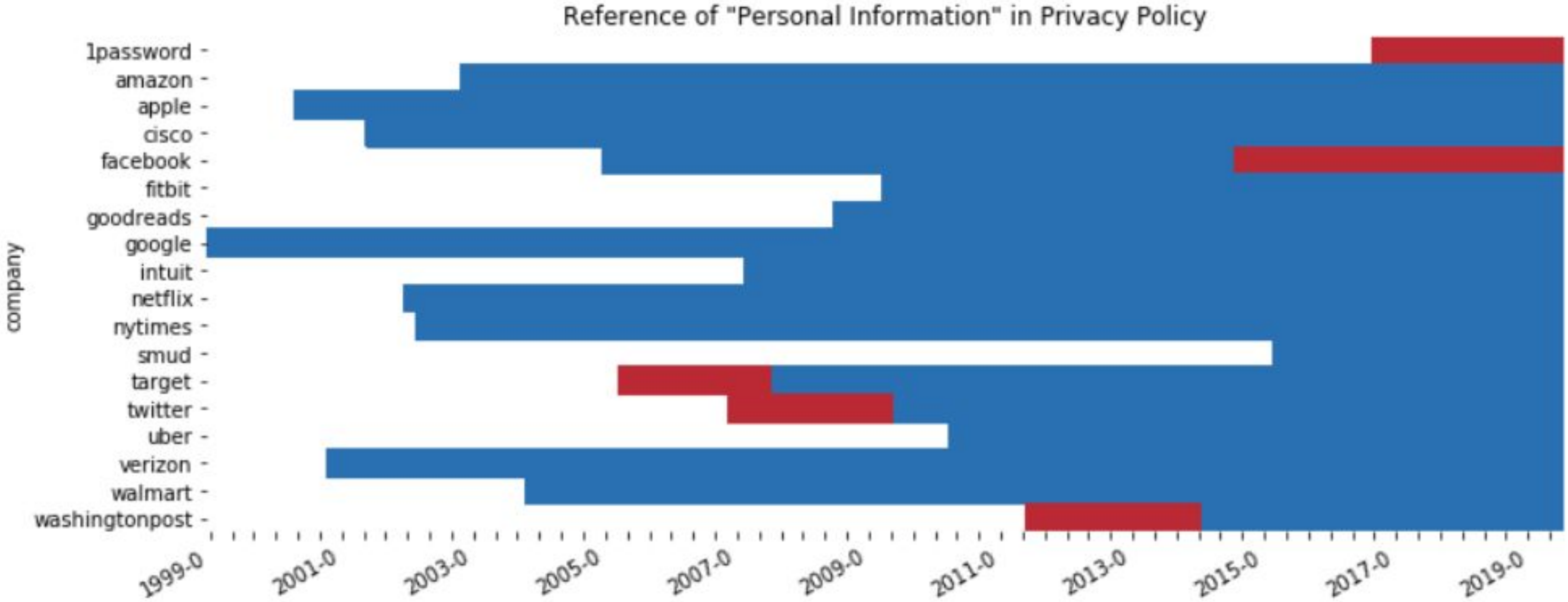
Other Datasets: https://usableprivacy.org/data

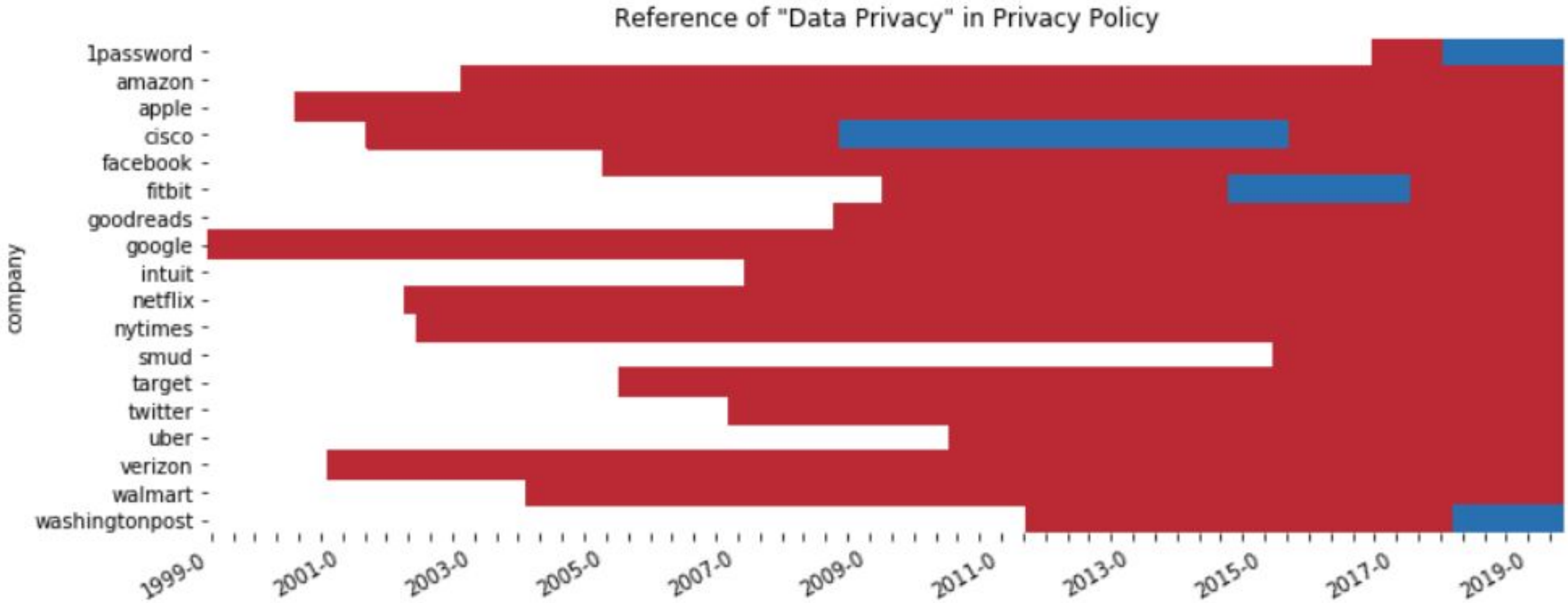
Github repo: https://github.com/workmanjack/privacy-policies-through-time

Appendix

Dataset Metadata




Reference of "Personal Information" in Privacy Policy
company
1password
amazon
apple
cisco
facebook
fitbit
goodreads
google
intuit
netflix
nytimes
smud
target
twitter
uber
verizon
walmart
washingtonpost
1999.0
2001.0
2003.0
2005.0
2007.0
2009.0
2011.0
2013.0
2015.0
2017.0
2019.0


Reference of "Data Privacy" in Privacy Policy
company
1password
amazon
apple
cisco
facebook
fitbit
goodreads
google
intuit
netflix
nytimes
smud
target
twitter
uber
verizon
walmart
washingtonpost
1999-0
2001-0
2003-0
2005-0
2007-0
2009-0
2011-0
2013-0
2015-0
2017-0
2019-0