

# *Mis-shapes, Mistakes, Misfits:* An Analysis of Domain Classification Services

Pelayo Vallina, Victor Le Pochat, Álvaro Feal,  
Marius Paraschiv, Julien Gamba, Tim Burke,  
Oliver Hohlfeld, Juan Tapiador, Narseo Vallina-Rodriguez





[NBA Draft](#)  
[Wimbledon](#)

[ONSALE](#)

**FIVE \$1,000**  
**Winners Click Now!**

[Holyfield  
vs. Tyson](#)

Search [options](#)

[Yellow Pages](#) - [People Search](#) - [Maps](#) - [Classifieds](#) - [News](#) - [Stock Quotes](#) - [Sports Scores](#)

- [Arts and Humanities](#)

[Architecture](#), [Photography](#), [Literature](#)...

- [Business and Economy \[Xtra!\]](#)

[Companies](#), [Investing](#), [Employment](#)...

- [Computers and Internet \[Xtra!\]](#)

[Internet](#), [WWW](#), [Software](#), [Multimedia](#)...

- [Education](#)

[Universities](#), [K-12](#), [College Entrance](#)...

- [Entertainment \[Xtra!\]](#)

[Cool Links](#), [Movies](#), [Music](#), [Humor](#)...

- [News and Media \[Xtra!\]](#)

[Current Events](#), [Magazines](#), [TV](#), [Newspapers](#)...

- [Recreation and Sports \[Xtra!\]](#)

[Sports](#), [Games](#), [Travel](#), [Autos](#), [Outdoors](#)...

- [Reference](#)

[Libraries](#), [Dictionaries](#), [Phone Numbers](#)...

- [Regional](#)

[Countries](#), [Regions](#), [U.S. States](#)...

- [Science](#)

[CS](#), [Biology](#), [Astronomy](#), [Engineering](#)...

**OpenDNS** / [dashboard](#)

[HOME](#)

[STATS](#)

[SETTINGS](#)

[MY ACCOUNT](#)

## Web Content Filtering

### Choose your filtering level

**High**

Protects against all adult-related sites, illegal activity, social networking sites, video sharing sites, and general time-wasters.

26 categories in this group - [View](#) - [Customize](#)

**Moderate**

Protects against all adult-related sites and illegal activity.

13 categories in this group - [View](#) - [Customize](#)

**Low**

Protects against pornography.

4 categories in this group - [View](#) - [Customize](#)

**None**

Nothing blocked.

**Custom**

Choose the categories you want to block.

Academic Fraud

Adult Themes

Adware

Alcohol

Anime/Manga/Webcomic

Auctions

Automotive

Blogs

Business Services

Chat

Classifieds

Dating

Drugs

Ecommerce/Shopping

Educational Institutions

File Storage

Financial Institutions

Forums/Message boards

Gambling

Games

German Youth Protection

# web directories (manually edited)

# classification engines (automated)

# Why does the quality of these services matter?

- › *End users*: incorrect categories affect reliability
  - » over/underblocking in content filtering
- › *Academia*: domain sample or results depend on them
  - » 2019 top conferences: 24 papers
  - » lack of trust → resort to manual classification

# Services are opaque on how they operate

The Forcepoint Master Database contains the industry's most accurate, current and comprehensive classification of URLs. We use proprietary classification software and

Webshrinker uses advanced Machine Learning algorithms.

***Collect the best websites for any topic!***

Validation? Training set? Comprehensiveness?

# **Methodology**

## **Empirical validation**

**Deep dive:  
human labeling  
& case studies**

**Discussion  
&  
Conclusion**

# Methodology

Empirical  
validation

Deep dive:  
human labeling  
& case studies

Discussion  
&  
Conclusion

Inputs	Outputs	Purpose	Updates	Access
--------	---------	---------	---------	--------



**Bitdefender®**



**OpenDNS**



Inputs	Outputs	Purpose	Updates	Access
--------	---------	---------	---------	--------



## Inputs

## Outputs

## Purpose

## Updates

## Access



**FortiGuard Labs**

Global threat research and response



**OpenDNS**



**VIRUSTOTAL**



**Alexa**



Inputs

Outputs

Purpose

Updates

Access

## Content filtering



**FortiGuard Labs**

Global threat research and response



webshrinker

**Marketing**

Purpose

Bitdefender®



**Forcepoint websense**

**OpenDNS**

## Threat assessment



**VIRUSTOTAL**



**Alexa**

**dmoz**

**Curlie** A small brown acorn icon next to the word 'Curlie'.

**Discovery**

Inputs

Outputs

Purpose

Updates

Access

(Mostly) automated



OpenDNS



Manual

## Inputs

## Outputs

## Purpose

## Updates

## Access



OpenDNS



# Methodology

Empirical  
validation

Deep dive:  
human labeling  
& case studies

Discussion  
&  
Conclusion

**Methodology**

**Empirical  
validation**

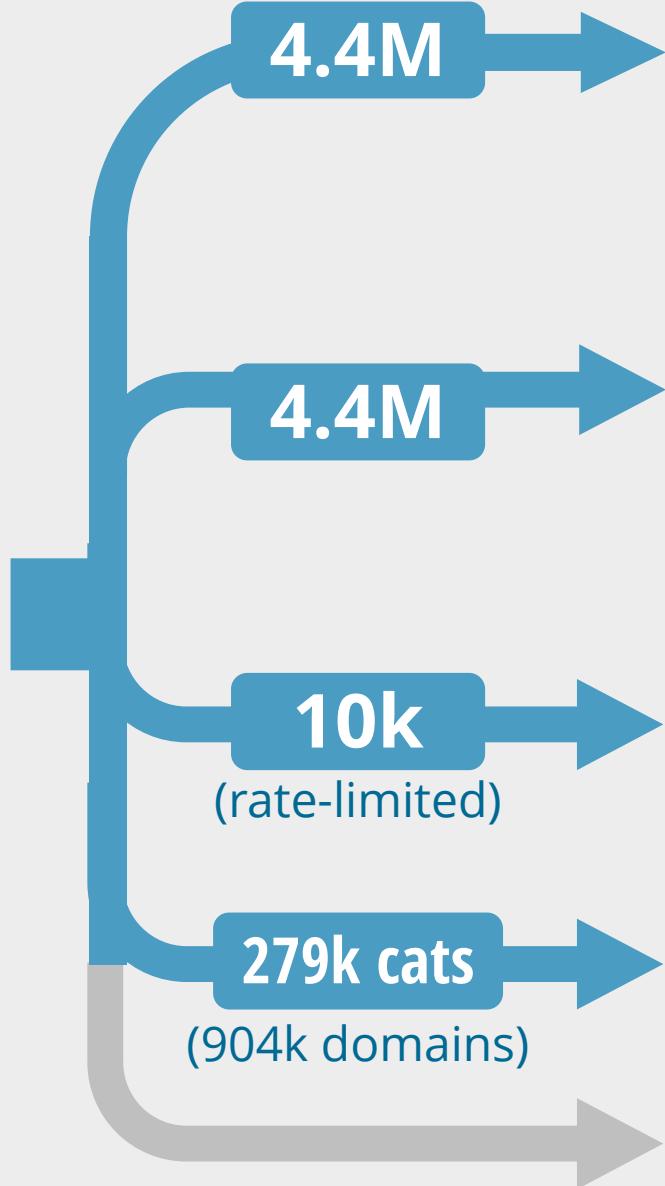
**Deep dive:  
human labeling  
& case studies**

**Discussion  
&  
Conclusion**

# Label gathering

 Alexa  
Top 1M domains  
  
Sept 1-30  
2019

Aggregate using Tranco  
↓  
**4.4M domains**



 FortiGuard Labs  
Global threat research and response

 McAfee™  OpenDNS

 VIRUSTOTAL

 Alexa Bitdefender Dr.WEB®  
Forcepoint websense TREND MICRO™

 Symantec  webshrinker  
 TREND MICRO™ direct

 Alexa direct

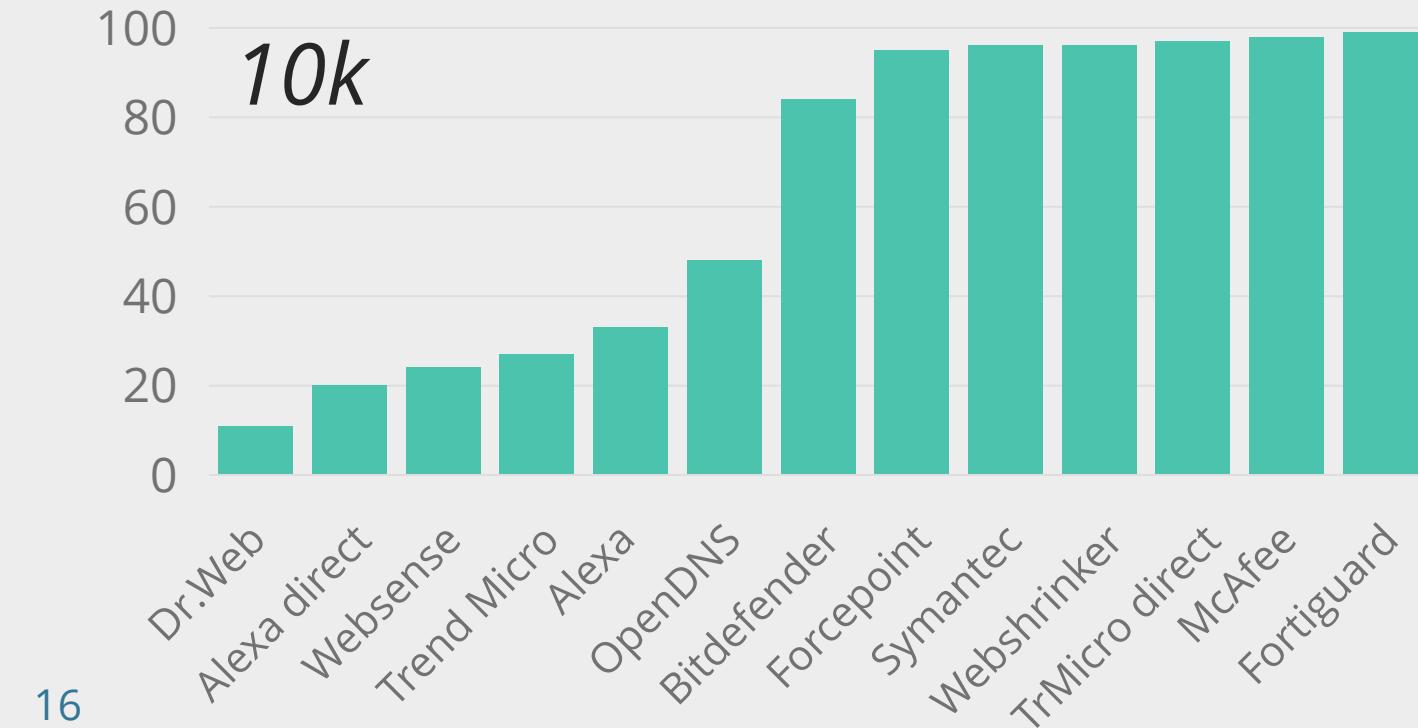
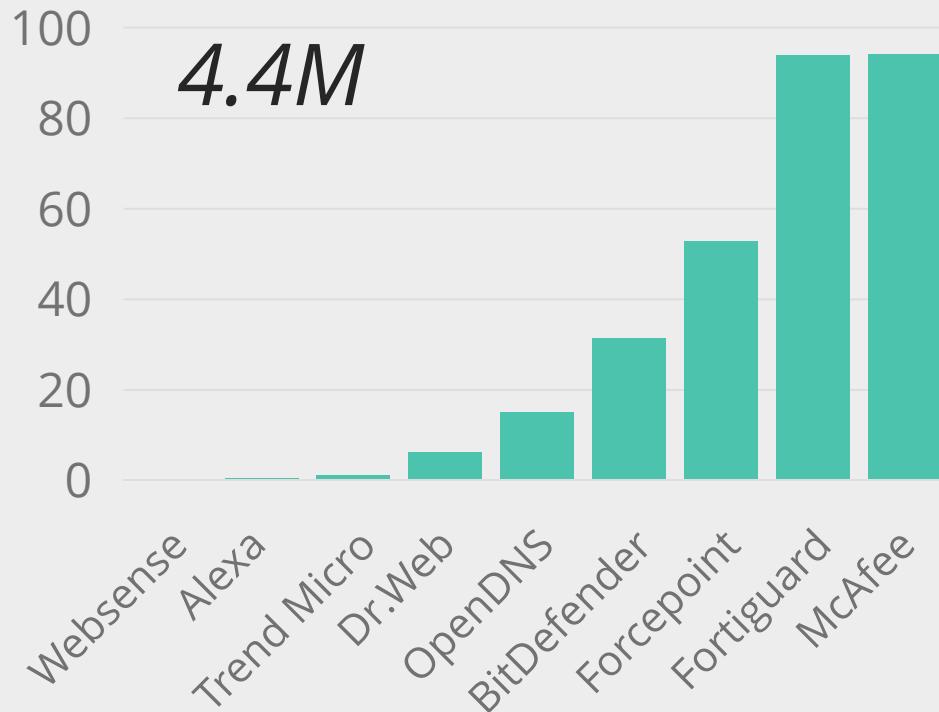
 Curlie  dmoz

# Service choice affects which domains are labeled

## › Coverage

- » ranges from <1% to 94%
- » is better for automated classification services

**Updates**



# Service choice affects which domains are labeled

- › Coverage
  - » ranges from <1% to 94%
  - » is better for automated classification services
- › Popular domains have better coverage
- › Subdomain coverage ranges from <1% to 99%
- › Inconsistent when directly sourced or through VirusTotal

Updates

Updates

Inputs

Access

# Service choice affects the taxonomy granularity

- › Security/content filtering: **fewer** categories
  - » As low as 12
  - » Easier setup
- › Marketing: **more** categories
  - » Up to 7.5k
  - » Fine-grained targeting

Purpose

# Service choice affects label interpretation

- › Inconsistencies between documented and observed labels
- › Multiple labels are uncommon
- › Subdomains inherit labels from parent
- › 3 out of 9 services updated labels
  - » Mostly for maliciousness

Access

Outputs

Inputs

Updates

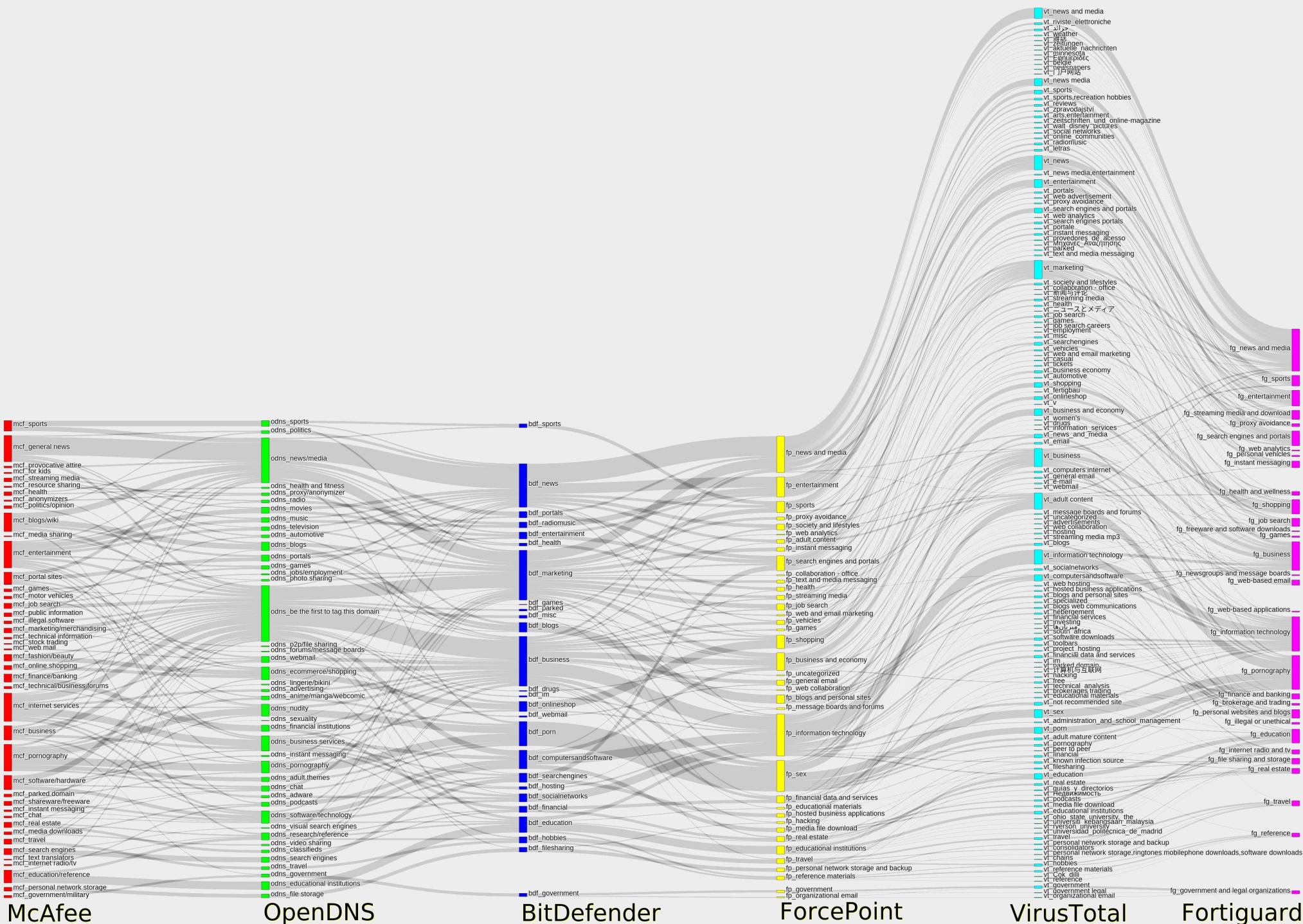
# Service choice affects label distribution

- › Disagreement
  - on distribution of labels over domains
    - » As measured through *mutual information*
- › Uneven distribution of labels over domains
  - » As measured through *label frequency*

Purpose

Updates

Purpose



**Methodology**

**Empirical  
validation**

**Deep dive:  
human labeling  
& case studies**

**Discussion  
&  
Conclusion**

## Methodology

### Empirical validation

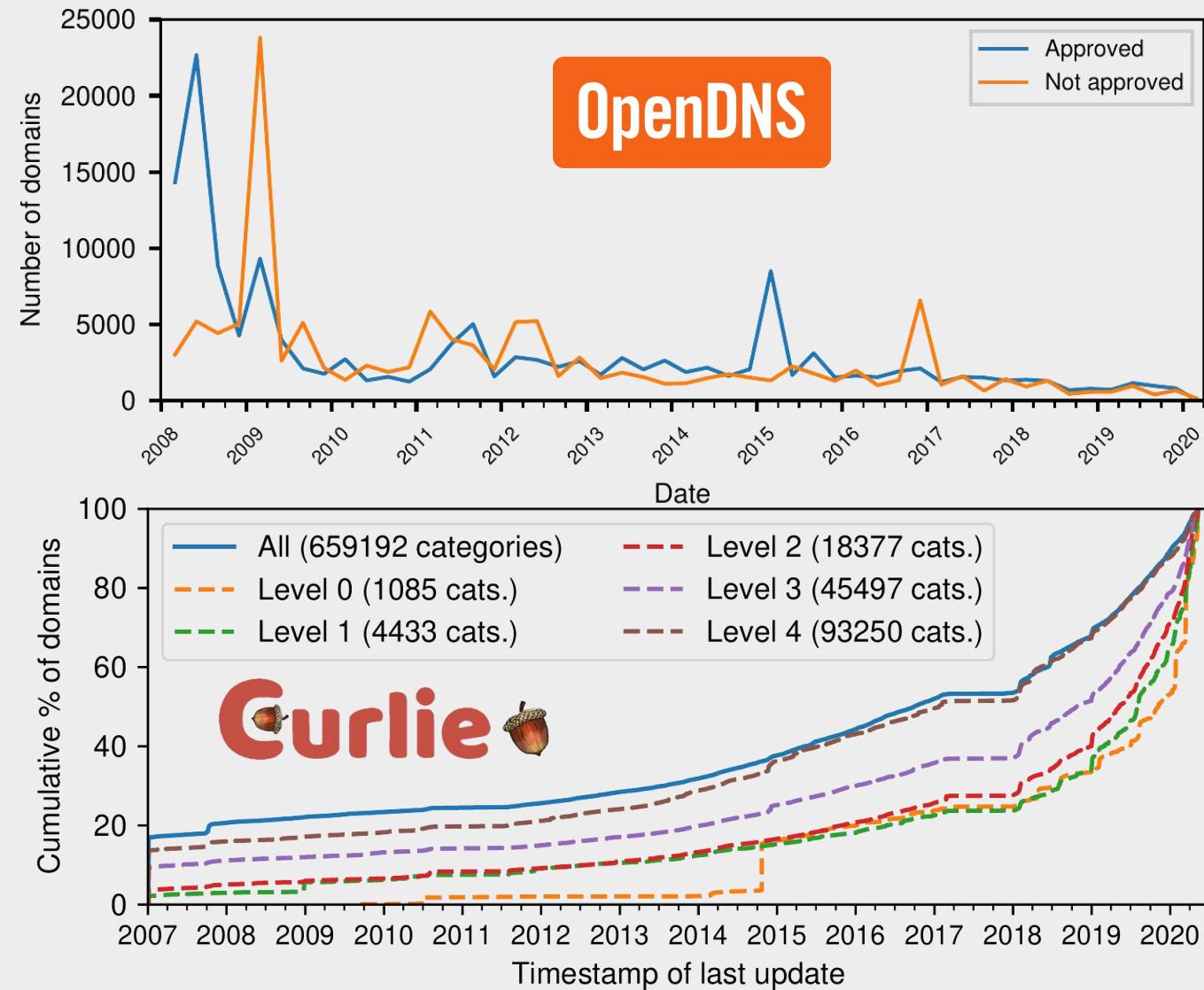
**Deep dive:  
human labeling  
& case studies**

**Discussion  
&  
Conclusion**

# Dynamics of human labeling may trigger biases

Participation concentrated

- › at *beginning* of project
  - » outdated labels?
- › with *few* users
  - » lack of peer review?
- › on *unlabeled* domains
  - » stale labels?

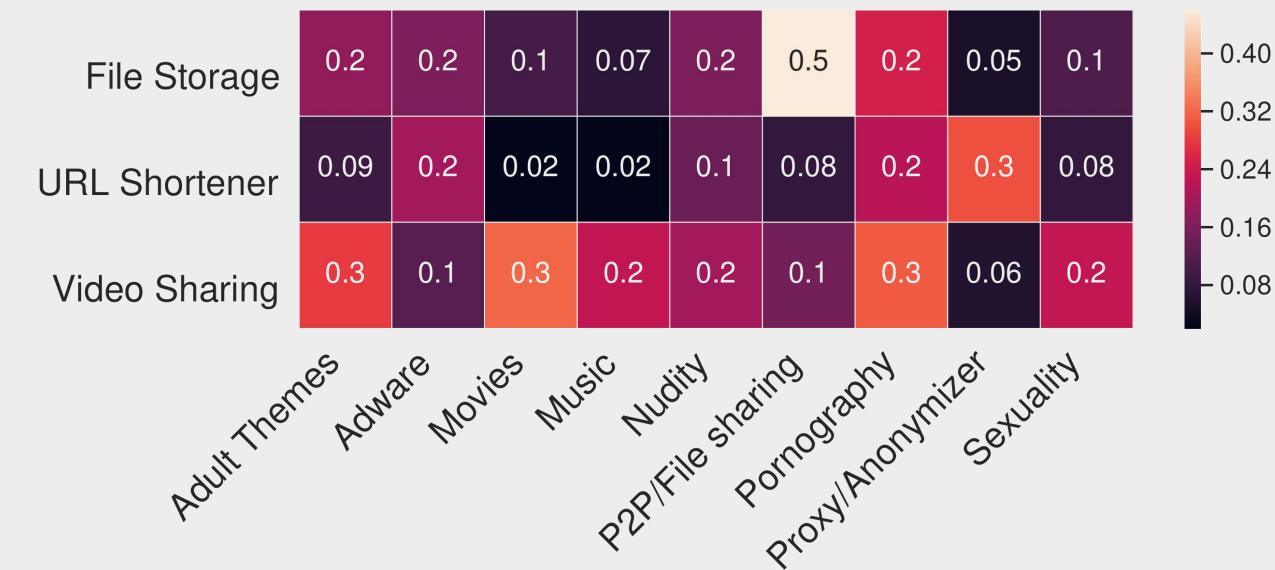


# *Disagreement* in human labeling may trigger biases

- › Label assignment is **not completely objective**

# *Disagreement* in human labeling may trigger biases

- › Label assignment is not completely objective
- › *Empirically:* Clusters of **correlated labels**



# *Disagreement* in human labeling may trigger biases

- › Label assignment is not completely objective
- › *Empirically:* Clusters of correlated labels
- › *Experimentally:* 35.5% **disagreement** among authors,  
71% matches community label

# We analyze services on *specialized use cases*

- › Intended usage → requirements → data source selection
- › Service selection → characteristics → coverage/accuracy
- › Estimate **suitability** for three case studies
  - » Obtain a manually curated list as “*ground truth*”
  - » Analyze *coverage* across domains
  - » Analyze *appropriateness* of labels

# Behavior differs widely for specialized use cases

- › Advertising and tracking
  - » *Curated list:* EasyList/EasyPrivacy
  - » *Finding:* few services label the domains **at all**, let alone as tracker
- › Adult content
  - » *Curated list:* [Val19] and gambling regulators
  - » *Finding:* 5 services label **correctly**, 3 others **hardly** label any
- › CDNs/hosting providers
  - » *Curated list:* signatures from WebPageTest
  - » *Finding:* **confusion** between service *function* and *content*

## Methodology

### Empirical validation

**Deep dive:  
human labeling  
& case studies**

**Discussion  
&  
Conclusion**

## Methodology

### Empirical validation

Deep dive:  
human labeling  
& case studies

Discussion  
&  
Conclusion

# Recommendations

- › We avoid recommending a *specific* service
  - » “Best” service depends on *use case* and *requirements*
  - » We cannot measure *semantic agreement* nor *correctness*
- › Our recommendations address *best practices*

# Recommendations

- › Coverage and accuracy may be **insufficient**
  - » Very service- and use case-dependent
  - » Consider **impact of errors**
- › Purpose and updates may introduce **biases**
  - » **Consult documentation** for taxonomy and label sources
  - » ... but **verify (and report)** manually, as **inconsistencies** exist
- › Taxonomies **differ** in size, scope and semantics
  - » Sound **aggregation** is **not obvious**

## Methodology

### Empirical validation

Deep dive:  
human labeling  
& case studies

Discussion  
&  
Conclusion

## Methodology

### Empirical validation

Deep dive:  
human labeling  
& case studies

Discussion  
&  
Conclusion

# *Mis-shapes, Mistakes, Misfits:* An Analysis of Domain Classification Services

Pelayo Vallina, Victor Le Pochat, Álvaro Feal,  
Marius Paraschiv, Julien Gamba, Tim Burke,  
Oliver Hohlfeld, Juan Tapiador, Narseo Vallina-Rodriguez



Universidad  
**Carlos III**  
de Madrid



Brandenburgische  
Technische Universität  
Cottbus - Senftenberg



# References

- › [Yan04] Hsin-Chang Yang and Chung-Hong Lee. 2004. A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications* 27, 4 (2004), 645–663. <https://doi.org/10.1016/j.eswa.2004.06.009>
- › [Qi09] Xiaoguang Qi and Brian D. Davison. 2009. Web Page Classification: Features and Algorithms. *Comput. Surveys* 41, 2, Article 12 (Feb. 2009), 31 pages. <https://doi.org/10.1145/1459352.1459357>
- › [Bru20] Renato Bruni and Gianpiero Bianchi. 2020. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications* 142, Article 113001 (2020), 14 pages. <https://doi.org/10.1016/j.eswa.2019.113001>
- › [Res04] Paul J. Resnick, Derek L. Hansen, and Caroline R. Richardson. 2004. Calculating Error Rates for Filtering Software. *Commun. ACM* 47, 9 (Sept. 2004), 67–71. <https://doi.org/10.1145/1015864.1015865>
- › [Ric02] Caroline R. Richardson, Paul J. Resnick, Derek L. Hansen, Holly A. Derry, and Victoria J. Rideout. 2002. Does Pornography-Blocking Software Block Access to Health Information on the Internet? *JAMA* 288, 22 (Dec. 2002), 2887–2894. <https://doi.org/10.1001/jama.288.22.2887>
- › [Sch18] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *IMC ’18*. 478–493. <https://doi.org/10.1145/3278532.3278574>
- › [LeP19] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *NDSS ’19*. 15. <https://doi.org/10.14722/ndss.2019.23386>
- › [Ahm20] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. 2020. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *WWW ’20*. 271–280. <https://doi.org/10.1145/3366423.3380113>
- › [Zeb20] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. 2020. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *WWW ’20*. 167–178. <https://doi.org/10.1145/3366423.3380104>
- › [Val19] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem. In *IMC ’19*.
- › [Seb16] Marcos Sebastián, Richard Rivera, Platon Kotzias, and Juan Caballero. 2016. AV-class: A Tool for Massive Malware Labeling. In *RAID ’16*. 230–253. [https://doi.org/10.1007/978-3-319-45719-2\\_11](https://doi.org/10.1007/978-3-319-45719-2_11)
- › [Lee13] Jung-Hyun Lee, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. 2013. Semantic Contextual Advertising Based on the Open Directory Project. *ACM Transactions on the Web* 7, 4, Article 24 (Nov. 2013), 22 pages. <https://doi.org/10.1145/2529995.2529997>
- › [Wei19] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You’ve Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *CCS ’19*. 149–166. <https://doi.org/10.1145/3319535.3363200>