Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment

https://dl.acm.org/doi/10.1145/3239574

JOHANNES KIESEL Bauhaus-Universität Weimar
FLORIAN KNEIST Ulm University
MILAD ALSHOMARY Paderborn University
BENNO STEIN Bauhaus-Universität Weimar
MATTHIAS HAGEN Martin-Luther-Universität Halle-Wittenberg
MARTIN POTTHAST Leipzig University

Presented by David Calano
CS 895 - Web Archiving Forensics
Old Dominion University
November 7th, 2022

Webis Web Archiver

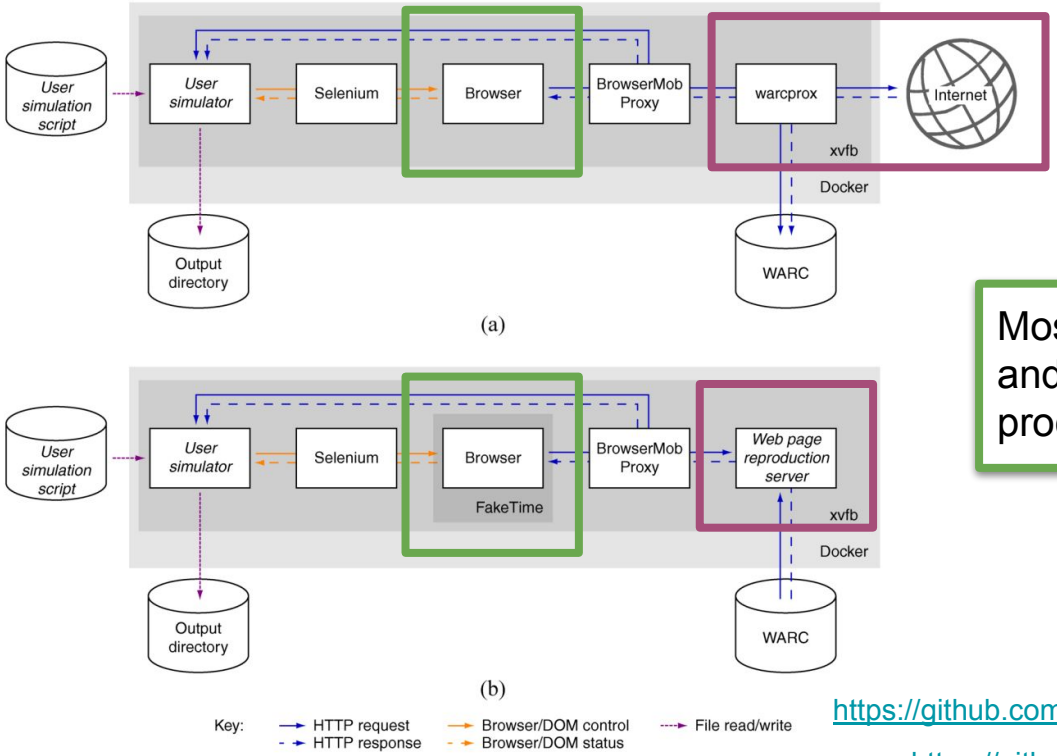


Figure 1

Mostly identical crawl and reproduction process

https://github.com/webis-de/webis-web-archiver

https://github.com/webis-de/scriptor

Breakdown of Dataset

Category	Sites	Pages	Most frequent site sub-categories	Most frequent page su
Adult	40	129	Computers (25), Image_Galleries (8), Society (2)	Computers (94), Image_ Society (3)
Arts	187	471	Television (59), Movies (19), Music (18)	Televi ion (152), Movie (50)
Business	229	489	Financial_Services (2), News_and_Media (24), Arts_and_Entertainment (22)	Financial_Services (116 News_and_Media (84), Arts_and_Entertainmen
Computers	444	1502	Internet (184), Software (143), Companies (60)	Internet (821), Software Companies (261)
Games	58	150	Video_Games (48), Gambling (6), Board_Games (1)	Video_Games (139), Ga Puzzles (2)
Health	28	36	Medicine (15), Conditions_and_Diseases (7), Nursing (5)	Medicine (21), Nursing Conditions_and_Disease
Home	42	88	Consumer_Information (19), Cooking (9), Personal_Finance (7)	Consumer_Information (17), Personal_Finance (15)
Kids_and_Teens	55	99	School_Time (21), Games (13), Computers (7)	School_Time (43), Gam Computers (10)
News	149	303	Newspapers (77), Breaking_News (13), Magazines_and_E-zines (8)	Newspapers (110), Head Weather (24)
Recreation	70	106	Travel (32), Autos (8), Humor (5)	Travel (63), Autos (10)
Reference	164	275	Education (105), Dictionaries (25), Libraries (21)	Education (128), Ask_a Dictionaries (47)
Regional	605	1182	North_America (395), Europe (112), Asia (78)	North_America (829), E Asia (144)
Science	74	116	Technology (19), Social_Sciences (16), Biology (12)	Technology (34), Social_Sciences (27), Biology (20)
Shopping	182	413	General_Merchandise (30), Clothing (26), Entertainment (16)	Entertainment (108), General_Merchandise (105), Auctions (51)
Society	100	139	Issues (20), Religion_and_Spirituality (15), Government (12)	Issues (27), Religion_and_Spirituality (18), Government (16)
Sports	73	123	Resources (18), Soccer (9), Baseball (6)	Resources (48), Soccer (16), Cricket (10)
World	1453	3437	Chinese_Simplified_CN (305), Russian (245), Deutsch (213)	Chinese_Simplified_CN (866), Russian (731), Deutsch (593)

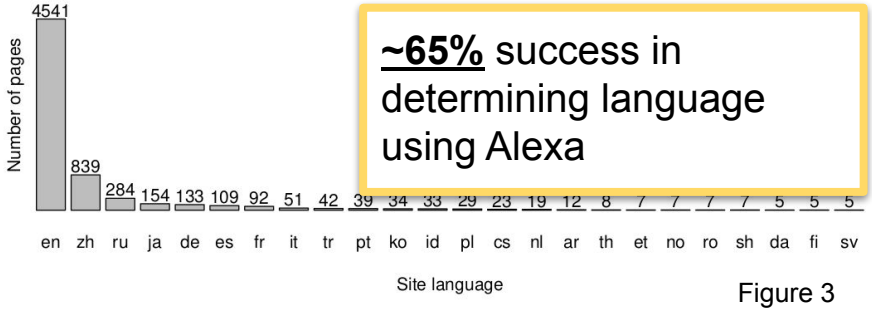
Table 1

https://alexa.com/topsites
https://webis.de/data/webis-web-archive-17.html

444 1502 Internet (184), Software (143), Companies (60)

605 1182 North_America (395), Europe (112), Asia (78)

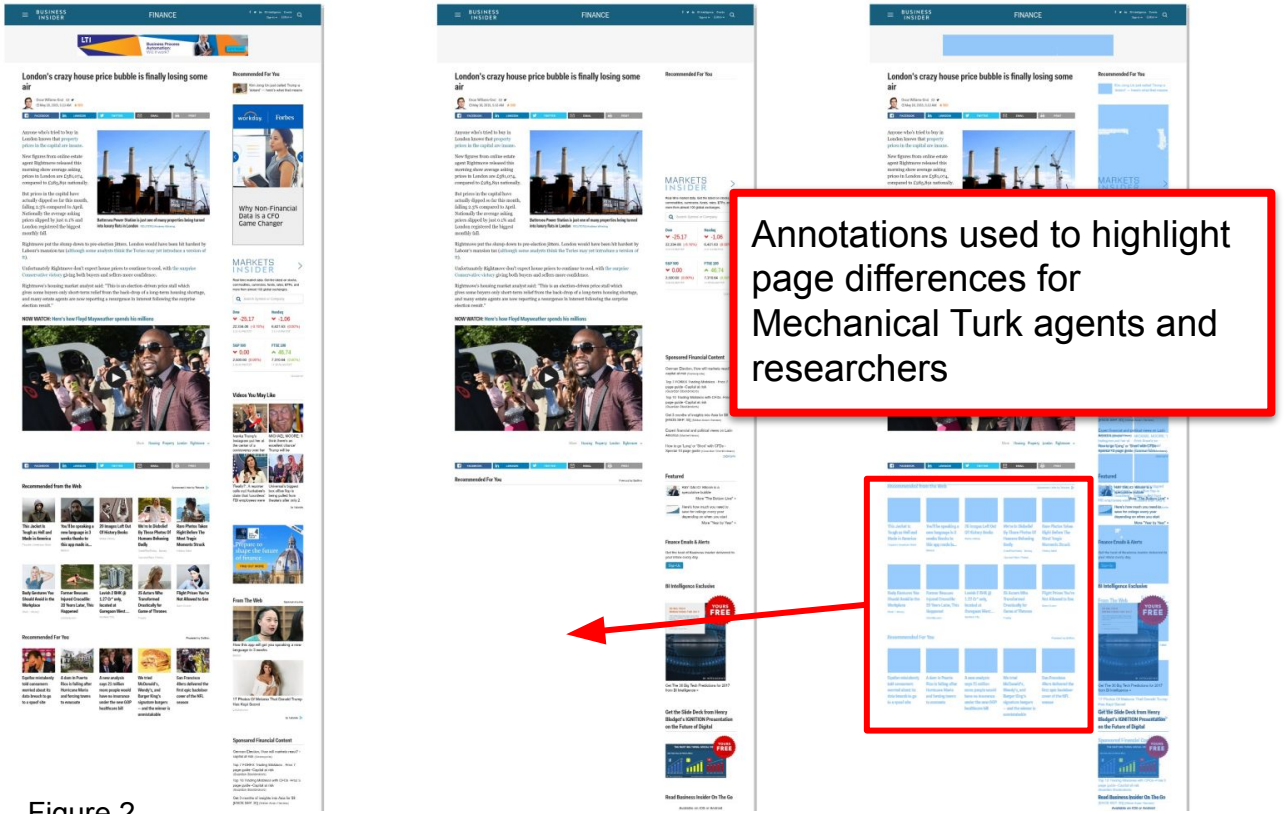
1453 3437 Chinese_Simplified_CN (305), Russian (245), Deutsch (213)



~65% success in determining language using Alexa

Figure 3

Web Capture vs Playback

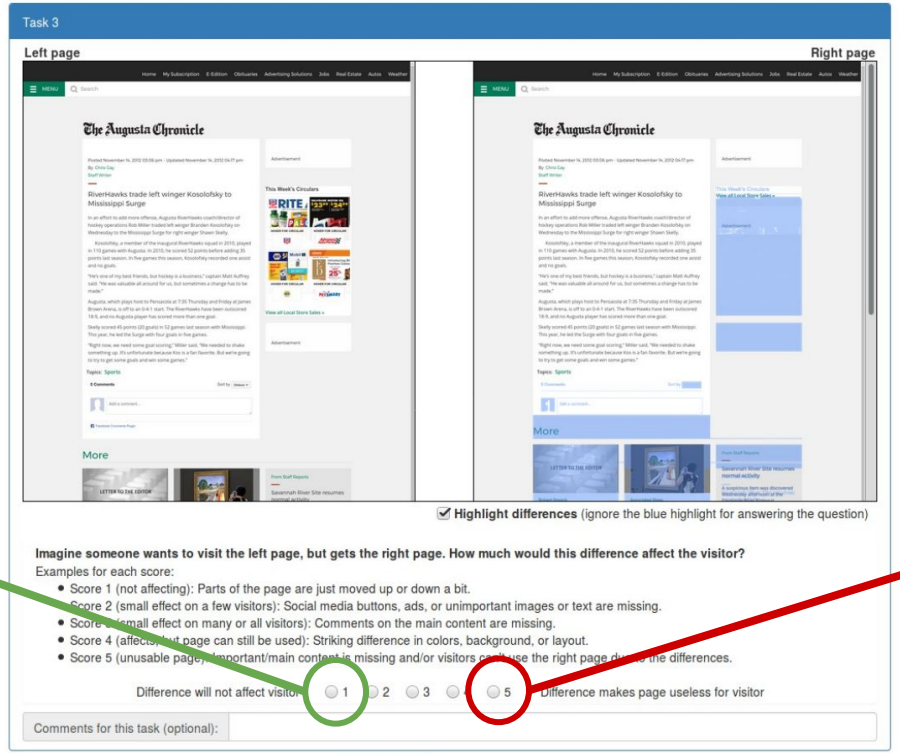


Annotations used to highlight page differences for Mechanical Turk agents and researchers

Figure 2

Mechanical Turk Used to Test Perceived Differences

1 for “best”

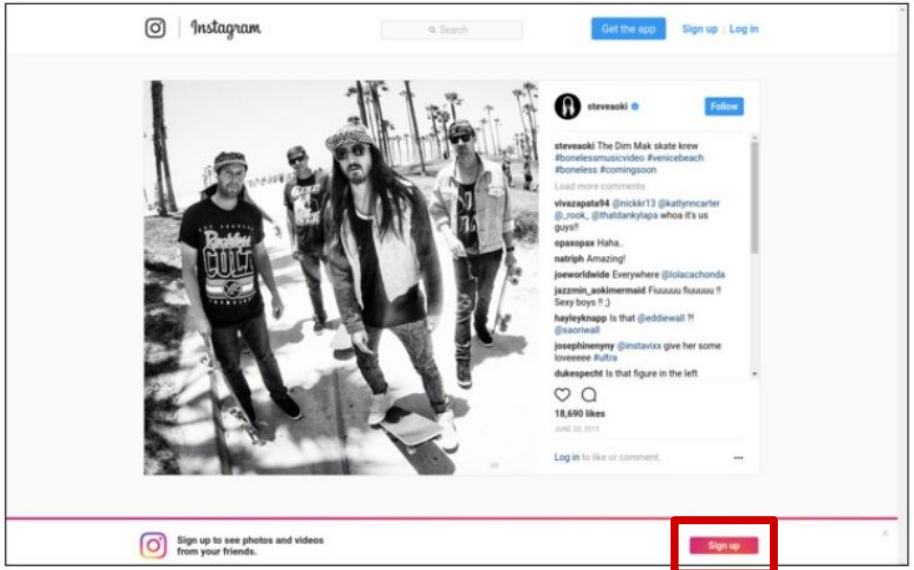


5 for “worst”

Figure 4

https://www.mturk.com/

1: Page Differences Have Negligible Effects



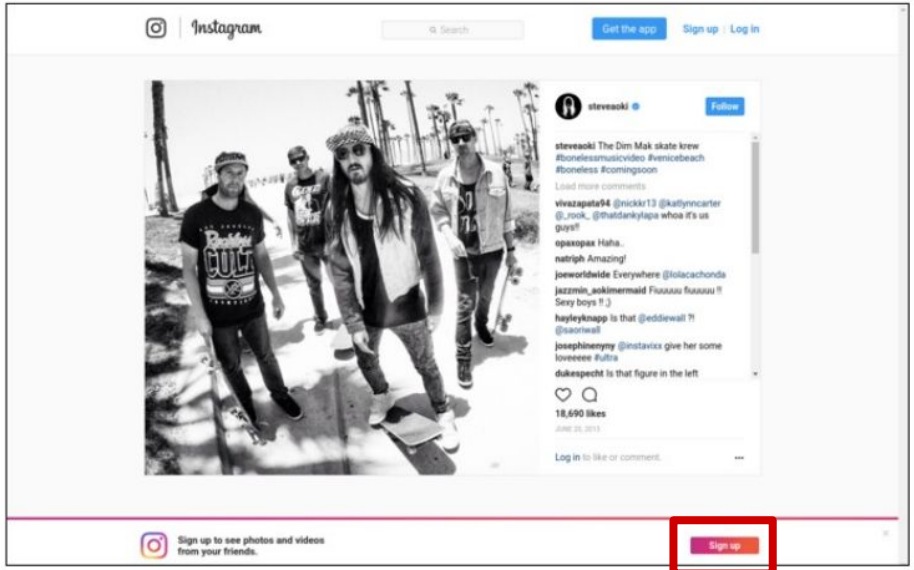


Figure 5

Slight difference in colors of animated button

2: Small Issues Affecting Small Numbers of People

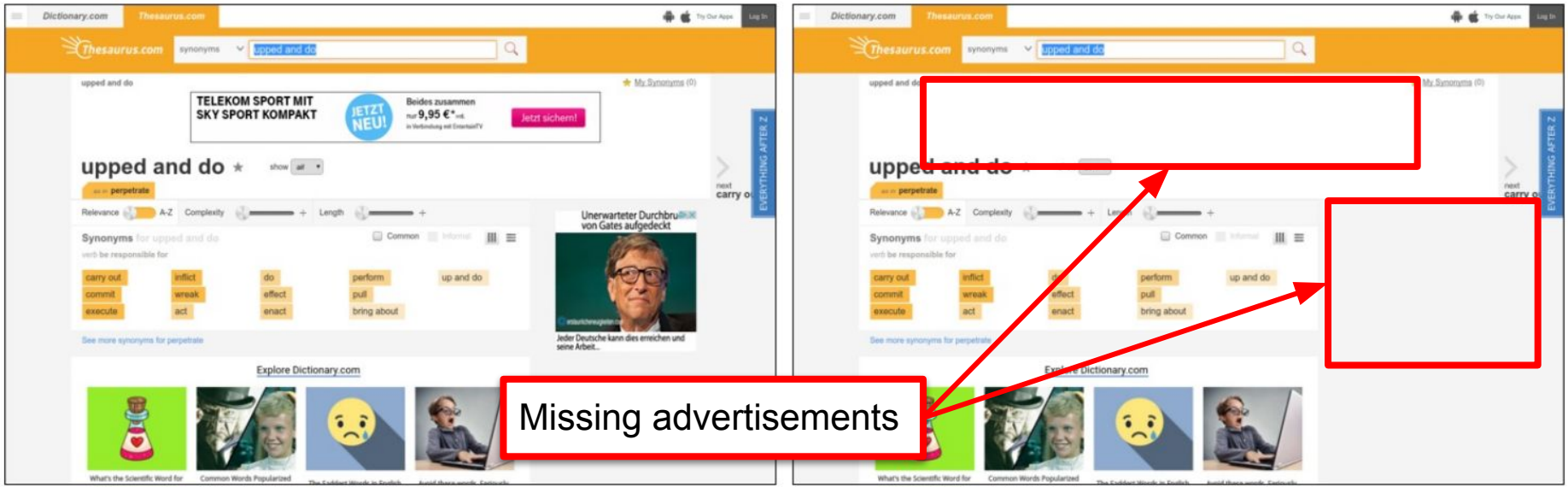


Figure 5

3: Page Has Small But Pervasive Differences

Figure 5

4: Page Is Usable Despite Significant Differences

Missing content not loaded at page initialization

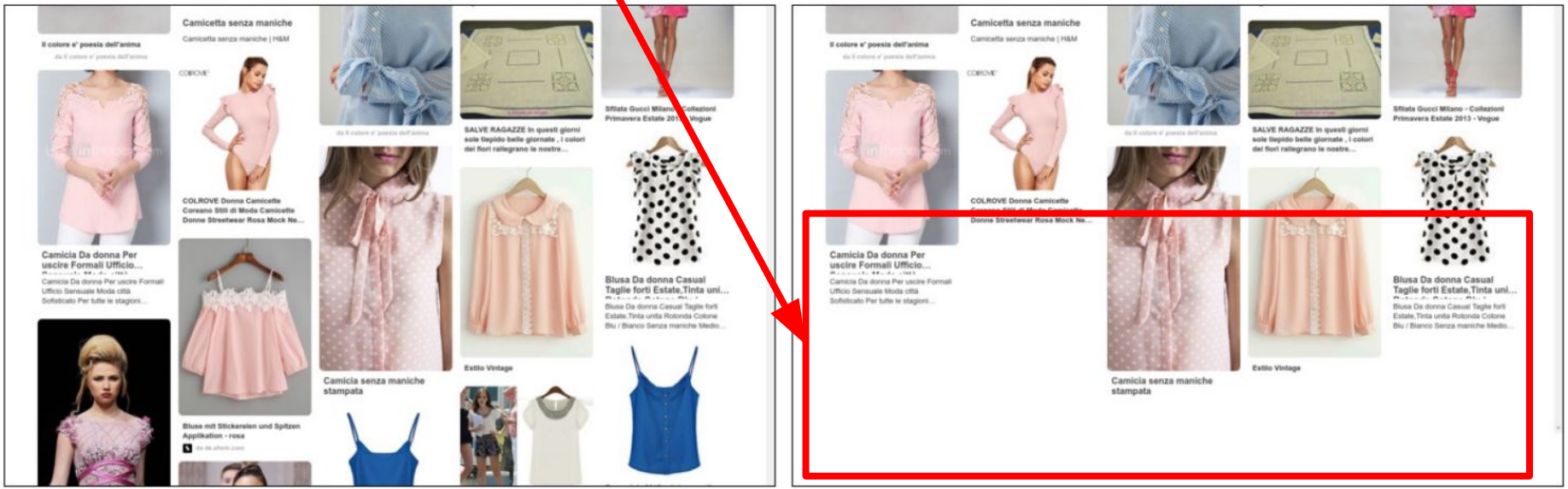


Figure 5

5: Page Is Effectively Unusable

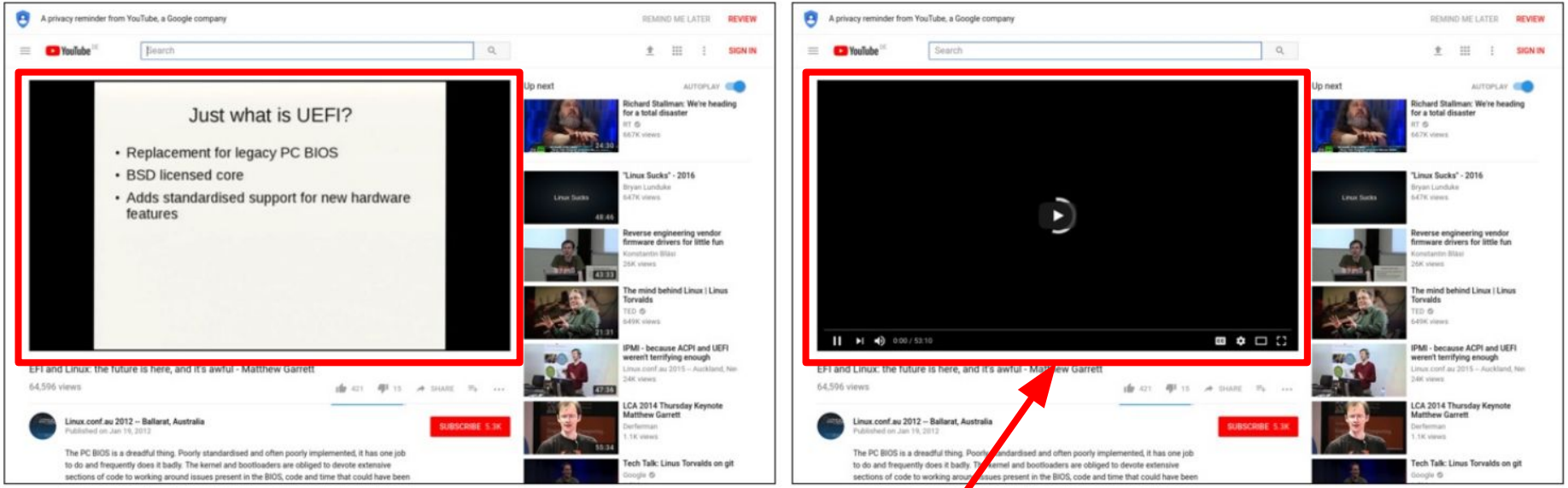


Figure 5

Primary content of page is missing or non-functional

Crowdsourced Score Breakdown

Reproductions	Score					Σ
	1	2	3	4	5	
Annotated	1942 (30.6%)	3307 (52.1%)	422 (6.7%)	318 (5.0%)	359 (5.7%)	6348
All	5594 (55.9%)	3307 (33.1%)	422 (4.2%)	318 (3.2%)	359 (3.6%)	10000

Table 2

Unannotated pages were reconfirmed by Mechanical Turk agents

Algorithm Comparison

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
https://en.wikipedia.org/wiki/Root-mean-square_deviation
http://justinfbrunelle.com/pubs/jcdl-2014-brunelle-damage.pdf
https://ws-dl.blogspot.com/2018/09/2018-09-03-lets-compare-memento-damage.html

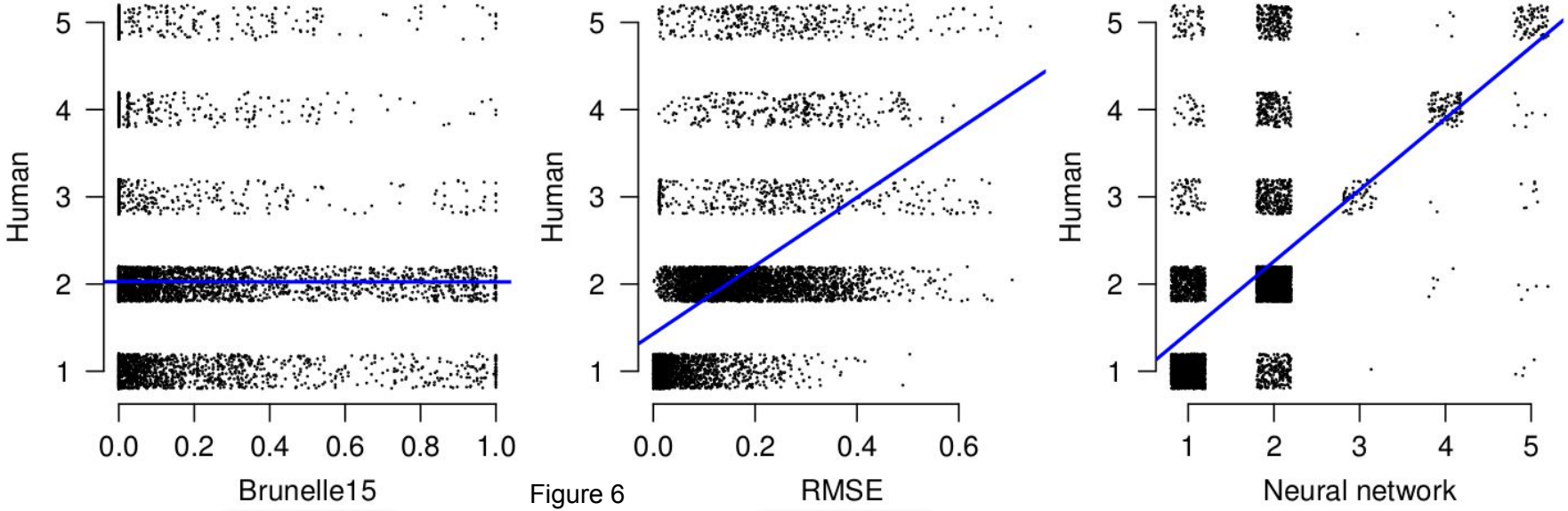


Figure 6

r = 0.0

r = 0.48

r = 0.57

Pearson Correlation Coefficient measures linear correlation between algorithmic results

Neural Network Accuracy

Min. Quality	Acc.	Prec.	Rec.	F_1
1	90.8%	94.1%	84.4%	89.0%
2	91.4%	94.4%	22.9%	36.9%
3	94.8%	88.1%	27.3%	41.7%
4	97.0%	76.0%	22.0%	34.1%

Table 3

Useful for automatically determining pages if were not faithfully reproduced

Most important metric for debugging

Calculations exclude trivial cases with no difference to screenshots

Comparison of Faithful Reproducibility for Replay Engines

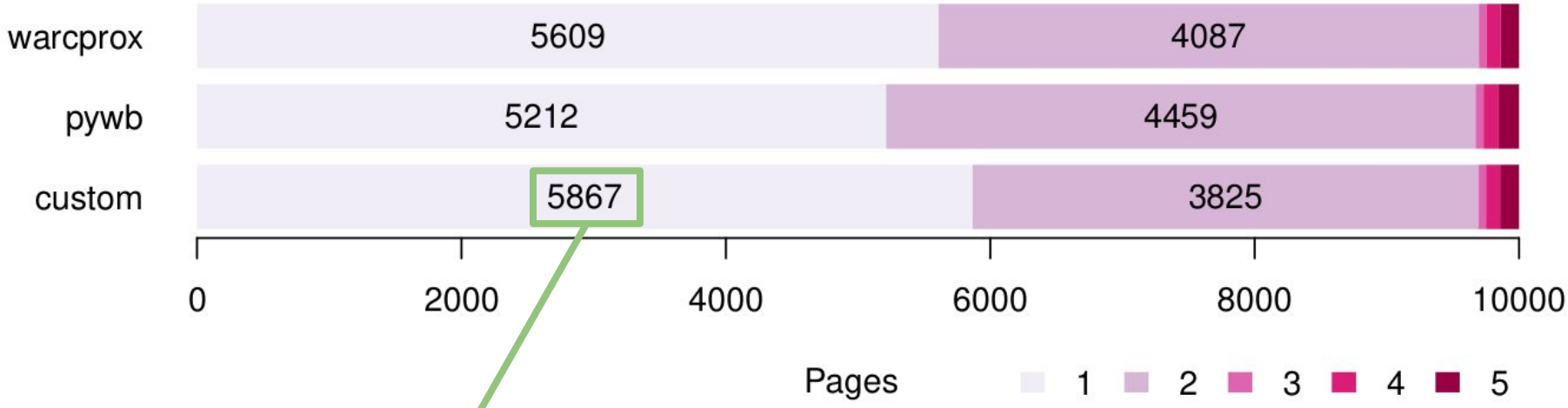


Figure 7

2.58% improvement over warcprox

6.55% improvement over pywb

https://github.com/internetarchive/warcprox

https://github.com/webrecorder/pywb

Conclusions

- Sample data set could be larger to improve accuracy and edge cases
- Using screenshot comparisons, neural networks can be trained to rapidly assess reproduced web page quality
- Move to automated calculation approaches over heuristics
- Check out Justin Brunelle's WS-DL blog post for more details: https://ws-dl.blogspot.com/2018/09/2018-09-03-lets-compare-memento-damage.html