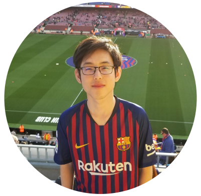
Jawa:

Web Archival in the Era of JavaScript



Ayush Goel
University of Michigan

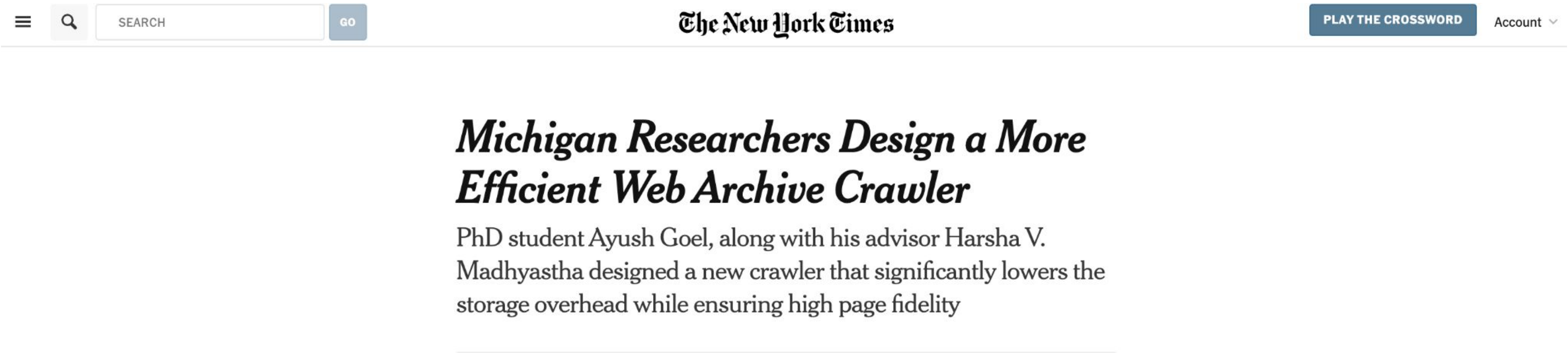
Jingyuan Zhu
University of Michigan

Ravi Netravali
Princeton University

Harsha V. Madhyastha
University of Michigan

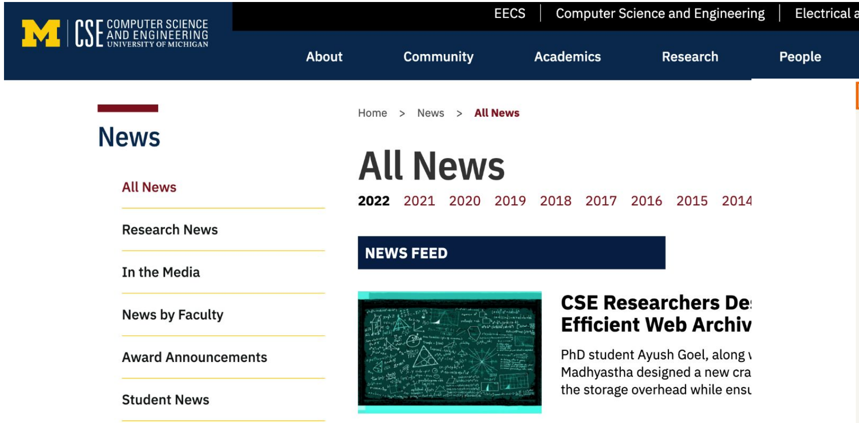
How to **reduce the storage overhead** of web archives and **improve the quality** of archived pages?

A Couple of Days After the Conference..



Michigan Researchers Design a More Efficient Web Archive Crawler

PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity



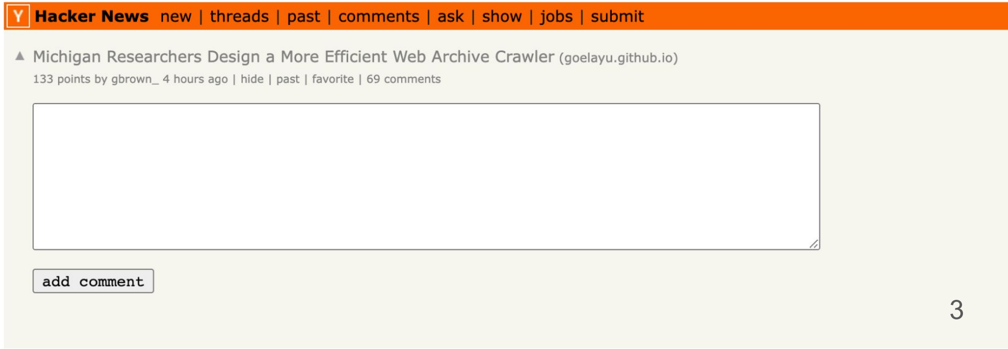
All News

2022 2021 2020 2019 2018 2017 2016 2015 2014

NEWS FEED

CSE Researchers De
Efficient Web Archiv

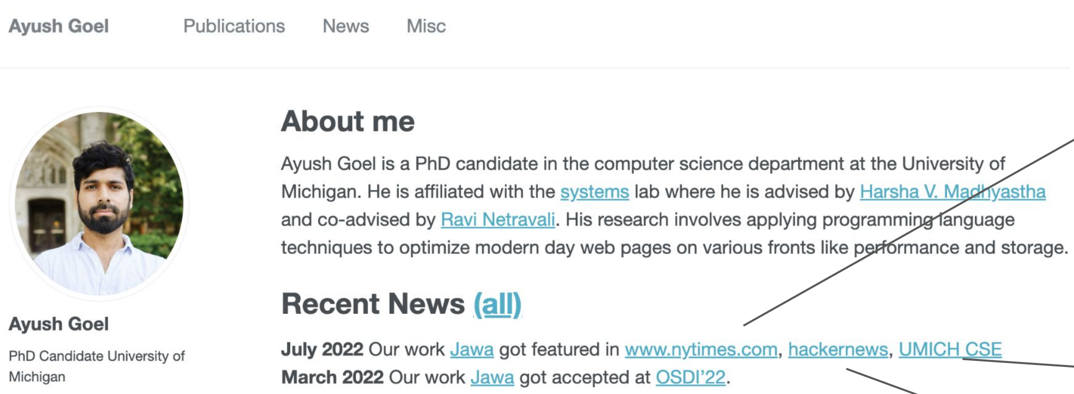
PhD student Ayush Goel, along
Madhyastha designed a new cra
the storage overhead while ensu


Y Hacker News new | threads | past | comments | ask | show | jobs | submit
add comment

▲ Michigan Researchers Design a More Efficient Web Archive Crawler (goelayu.github.io)

133 points by gbrown_ 4 hours ago | hide | past | favorite | 69 comments

A Couple of Days After the Conference..



About me

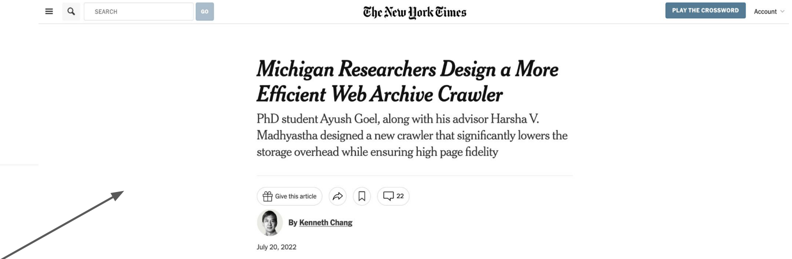
Ayush Goel is a PhD candidate in the computer science department at the University of Michigan. He is affiliated with the systems lab where he is advised by Harsha V. Madhyastha and co-advised by Ravi Netravali. His research involves applying programming language techniques to optimize modern day web pages on various fronts like performance and storage.

Ayush Goel

PhD Candidate University of Michigan

Recent News (all)

July 2022 Our work Jawa got featured in www.nytimes.com, hackernews, UMICH CSE
March 2022 Our work Jawa got accepted at OSDI'22.

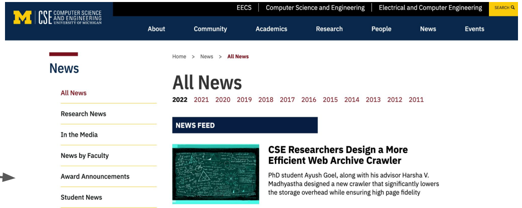


Michigan Researchers Design a More Efficient Web Archive Crawler

PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity

By Kenneth Chang

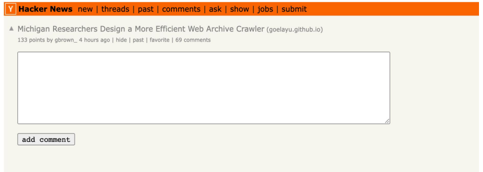
July 20, 2022



All News

CSE Researchers Design a More Efficient Web Archive Crawler

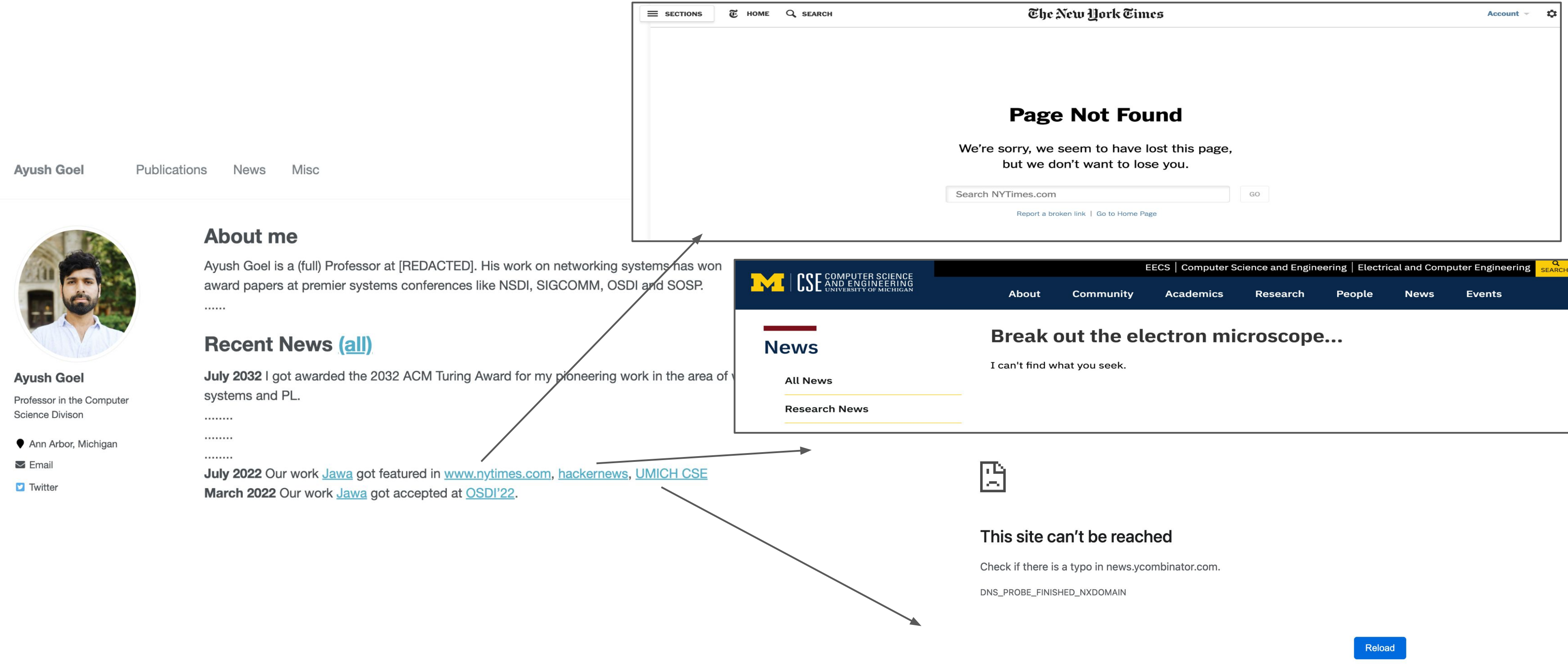
PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity


Y
Hacker News new | threads | past | comments | ask | show | jobs | submit
add comment

▲ Michigan Researchers Design a More Efficient Web Archive Crawler (goelayu.github.io)

133 points by gbrown_ 4 hours ago | hide | past | favorite | 69 comments

10 Years From Now (2032)....



Ayush Goel Publications News Misc

About me

Ayush Goel is a (full) Professor at [REDACTED]. His work on networking systems has won award papers at premier systems conferences like NSDI, SIGCOMM, OSDI and SOSP.

......

Recent News (all)

July 2032 I got awarded the 2032 ACM Turing Award for my pioneering work in the area of systems and PL.

....

....

....

July 2022 Our work Jawa got featured in www.nytimes.com, hackernews, UMICH CSE

March 2022 Our work Jawa got accepted at OSDI'22.

Ayush Goel

Professor in the Computer Science Divison

Ann Arbor, Michigan

Email

Twitter

Page Not Found

We're sorry, we seem to have lost this page, but we don't want to lose you.

EECS | Computer Science and Engineering | Electrical and Computer Engineering
SEARCH
CSE COMPUTER SCIENCE AND ENGINEERING UNIVERSITY OF MICHIGAN
About
Community
Academics
Research
People
News
Events
News
All News
Research News

Break out the electron microscope...

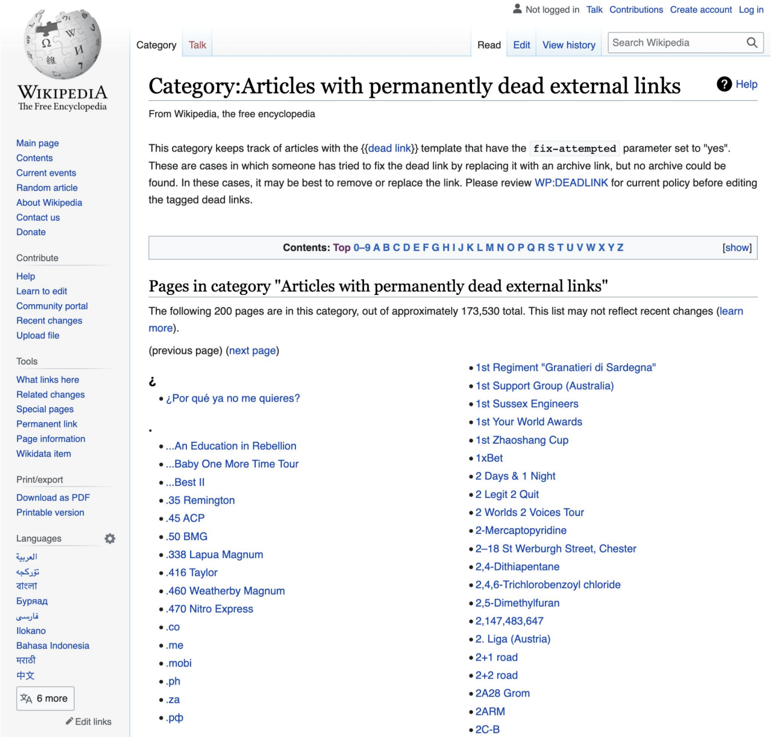
I can't find what you seek.

This site can't be reached

Check if there is a typo in news.ycombinator.com.

DNS_PROBE_FINISHED_NXDOMAIN

High Prevalence of Link Rot on the Web



Category:Articles with permanently dead external links

From Wikipedia, the free encyclopedia

This category keeps track of articles with the {{dead link}} template that have the `fix-attempted` parameter set to "yes". These are cases in which someone has tried to fix the dead link by replacing it with an archive link, but no archive could be found. In these cases, it may be best to remove or replace the link. Please review WP:DEADLINK for current policy before editing the tagged dead links.

Contents: Top 0–9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [show]

Pages in category "Articles with permanently dead external links"

The following 200 pages are in this category, out of approximately 173,530 total. This list may not reflect recent changes (learn more).

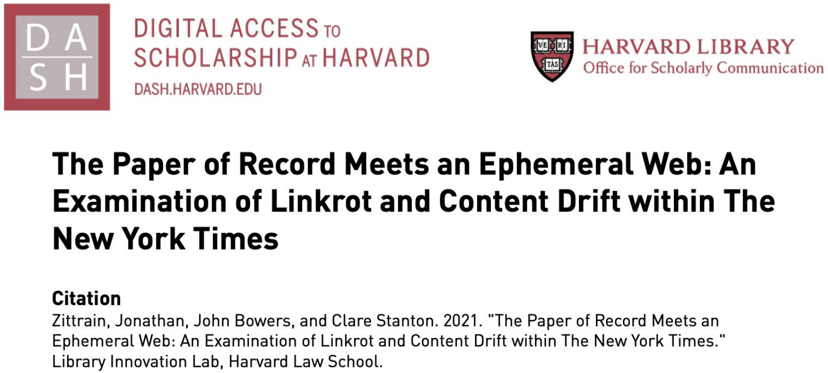
(previous page) (next page)

¿

- ¿Por qué ya no me quieres?

.

- ...An Education in Rebellion
- ...Baby One More Time Tour
- ...Best II
- .35 Remington
- .45 ACP
- .50 BMG
- .338 Lapua Magnum
- .416 Taylor
- .460 Weatherby Magnum
- .470 Nitro Express
- .co
- .me
- .mobi
- .ph
- .za
- .рф
- 1st Regiment "Granatieri di Sardegna"
- 1st Support Group (Australia)
- 1st Sussex Engineers
- 1st Your World Awards
- 1st Zhaoshang Cup
- 1xBet
- 2 Days & 1 Night
- 2 Legit 2 Quit
- 2 Worlds 2 Voices Tour
- 2-Mercaptopyridine
- 2–18 St Werburgh Street, Chester
- 2,4-Dithiapentane
- 2,4,6-Trichlorobenzoyl chloride
- 2,5-Dimethylfuran
- 2,147,483,647
- 2. Liga (Austria)
- 2+1 road
- 2+2 road
- 2A28 Grom
- 2ARM
- 2C-B

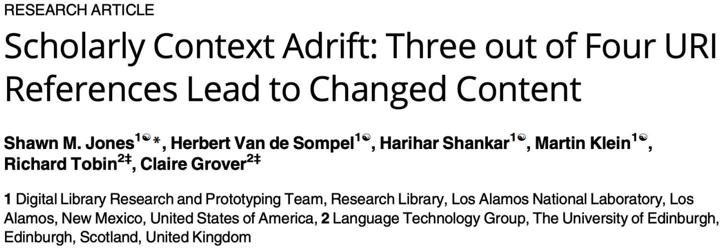


D A
S H
DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU
HARVARD LIBRARY
Office for Scholarly Communication

The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within The New York Times

Citation

Zittrain, Jonathan, John Bowers, and Clare Stanton. 2021. "The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within The New York Times." Library Innovation Lab, Harvard Law School.



RESEARCH ARTICLE

Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content

Shawn M. Jones[1]*, Herbert Van de Sompel[1], Harihar Shankar[1], Martin Klein[1], Richard Tobin[2‡], Claire Grover[2‡]

1 Digital Library Research and Prototyping Team, Research Library, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, **2** Language Technology Group, The University of Edinburgh, Edinburgh, Scotland, United Kingdom

Web Archives to the Rescue

Web Archives to the Rescue

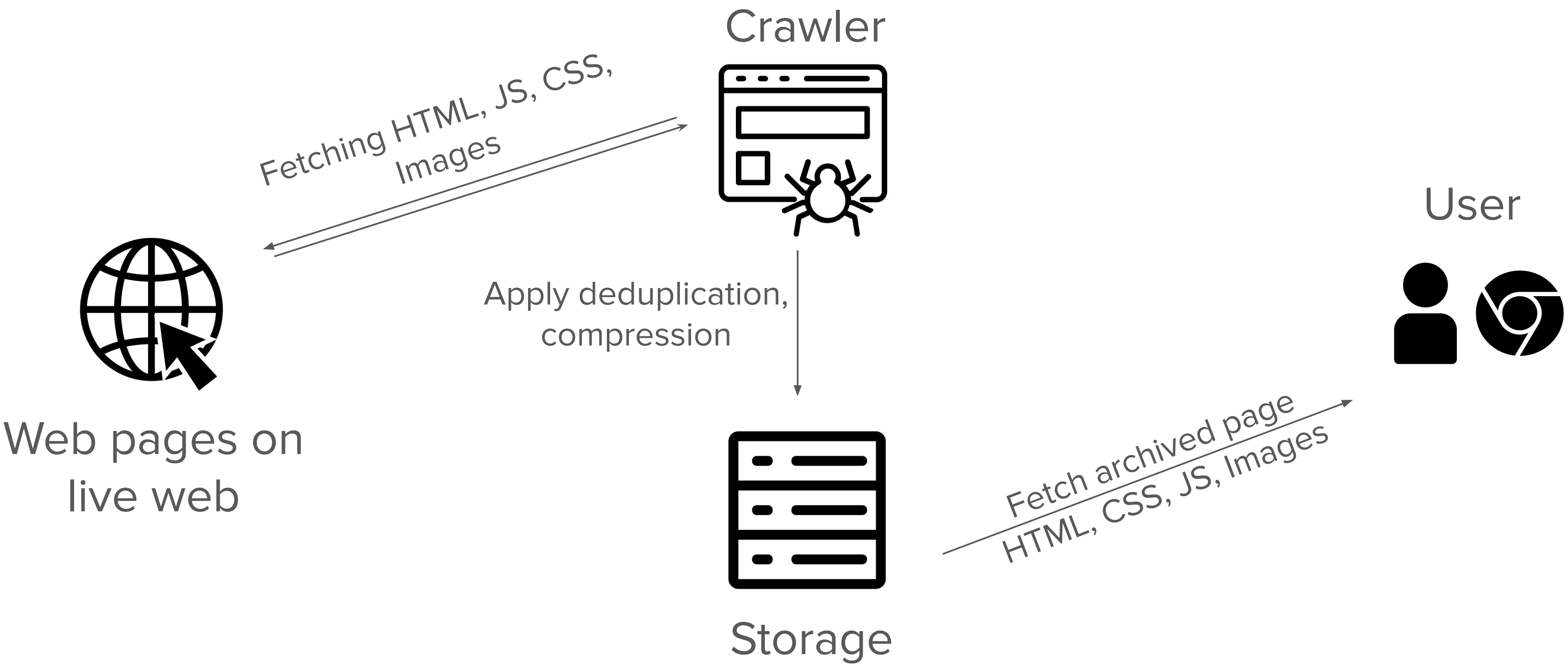
~700 Billion pages, ~100 Petabytes

LIBRARY
LIBRARY OF CONGRESS

UKWA
UK WEB ARCHIVE

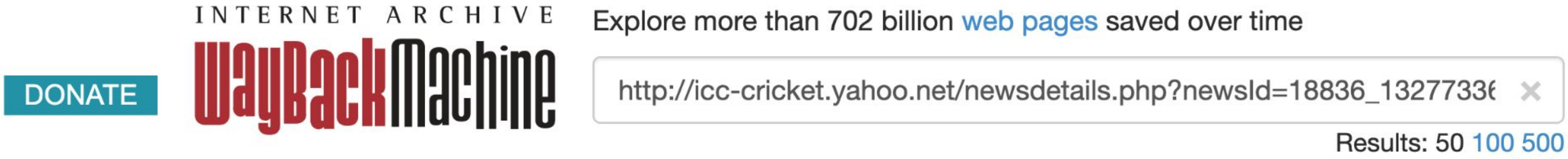
ARQUIVO.PT

How Modern Web Archives Operate?



Crawler
Fetching HTML, JS, CSS, Images
User
Apply deduplication, compression
Web pages on live web
Fetch archived page HTML, CSS, JS, Images
Storage

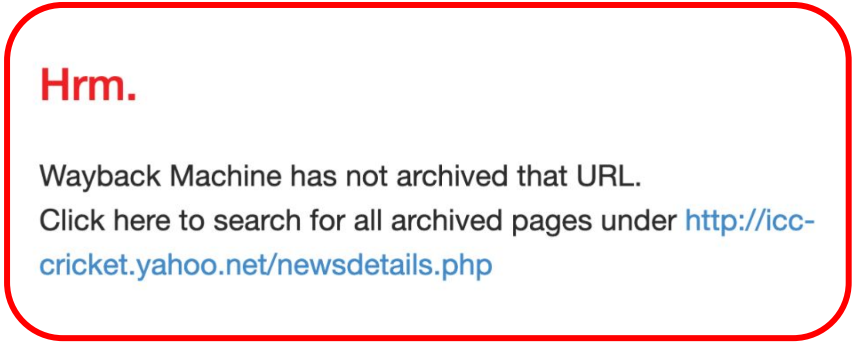
Problems with Web Archives: **No Snapshots for Many Pages**



Explore more than 702 billion web pages saved over time

http://icc-cricket.yahoo.net/newsdetails.php?newsId=18836_1327733(

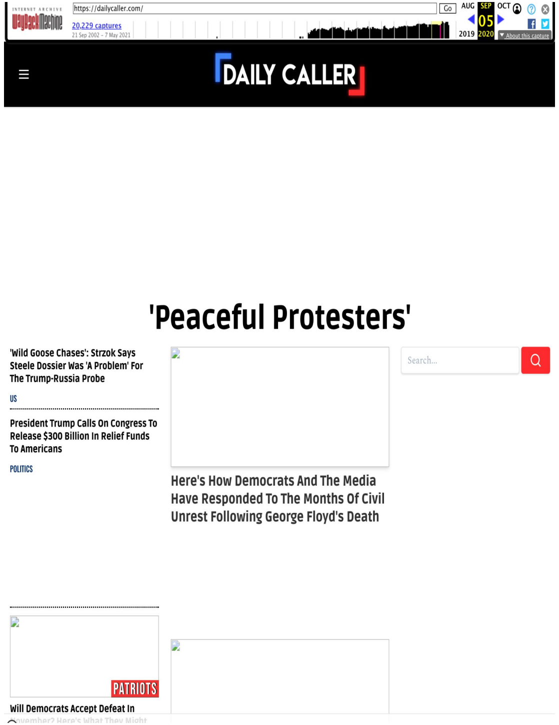
Results: 50 100 500



Hrm.

Wayback Machine has not archived that URL.
Click here to search for all archived pages under http://icc-cricket.yahoo.net/newsdetails.php

Problems with Web Archives: Poor Page Fidelity



INTERNET ARCHIVE WaybackMachine
https://dailycaller.com/ Go
20,229 captures
21 Sep 2002 – 7 May 2021
AUG SEP OCT
05
2019 2020
About this capture

DAILY CALLER

'Peaceful Protesters'

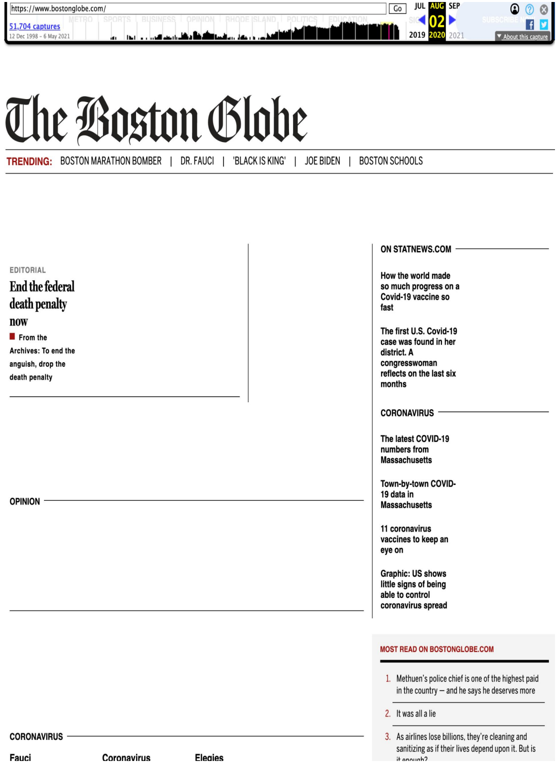
'Wild Goose Chases': Strzok Says Steele Dossier Was 'A Problem' For The Trump-Russia Probe
US

President Trump Calls On Congress To Release $300 Billion In Relief Funds To Americans
POLITICS

Here's How Democrats And The Media Have Responded To The Months Of Civil Unrest Following George Floyd's Death

Search...

PATRIOTS
Will Democrats Accept Defeat In November? Here's What They Might



https://www.bostonglobe.com/ Go
51,704 captures
12 Dec 1998 – 6 May 2021
JUL AUG SEP
02
2019 2020 2021
About this capture

The Boston Globe

TRENDING: BOSTON MARATHON BOMBER | DR. FAUCI | 'BLACK IS KING' | JOE BIDEN | BOSTON SCHOOLS

EDITORIAL
End the federal death penalty now
From the Archives: To end the anguish, drop the death penalty

OPINION

CORONAVIRUS
Fauci Coronavirus Elegies

ON STATNEWS.COM
How the world made so much progress on a Covid-19 vaccine so fast
The first U.S. Covid-19 case was found in her district. A congresswoman reflects on the last six months

CORONAVIRUS
The latest COVID-19 numbers from Massachusetts
Town-by-town COVID-19 data in Massachusetts
11 coronavirus vaccines to keep an eye on
Graphic: US shows little signs of being able to control coronavirus spread

MOST READ ON BOSTONGLOBE.COM
1. Methuen's police chief is one of the highest paid in the country — and he says he deserves more
2. It was all a lie
3. As airlines lose billions, they're cleaning and sanitizing as if their lives depend upon it. But is it enough?

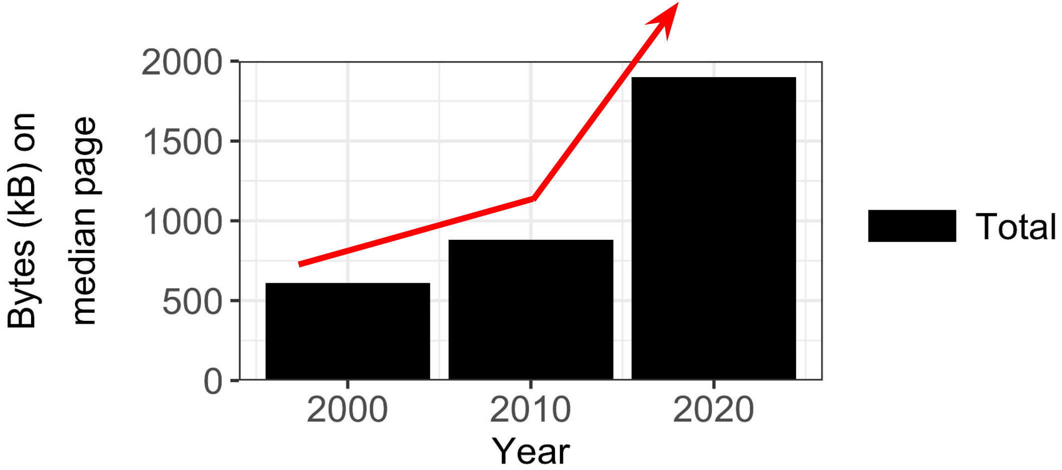
Outline

- ★ **Challenges faced by web archives**
- ★ Understanding the root cause
- ★ Our insights and design of a new crawler
- ★ Evaluation

Outline

- ★ Challenges faced by web archives
- ★ **Understanding the root cause**
- ★ Our insights and design of a new crawler
- ★ Evaluation

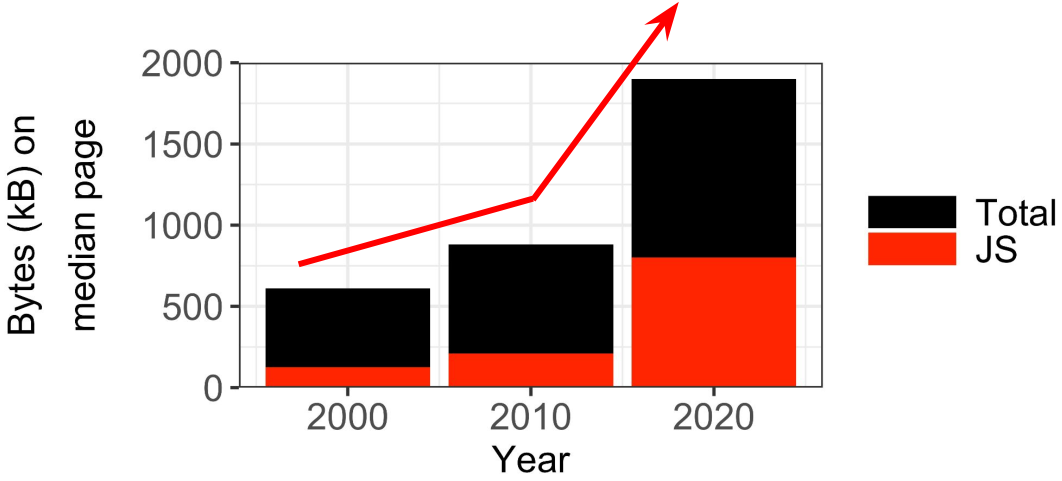
Root Cause: **Increasing JavaScript** on Web Pages



Corpus: Landing pages of 300 Alexa sites

Root Cause: **Increasing JavaScript** on Web Pages

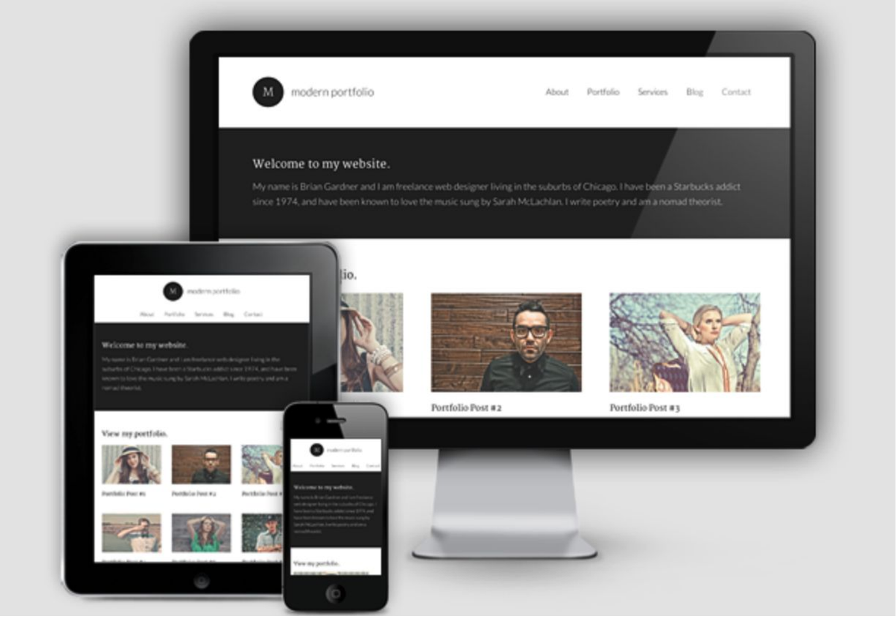
Per page snapshot **expensive** ➔ **Fewer snapshots**

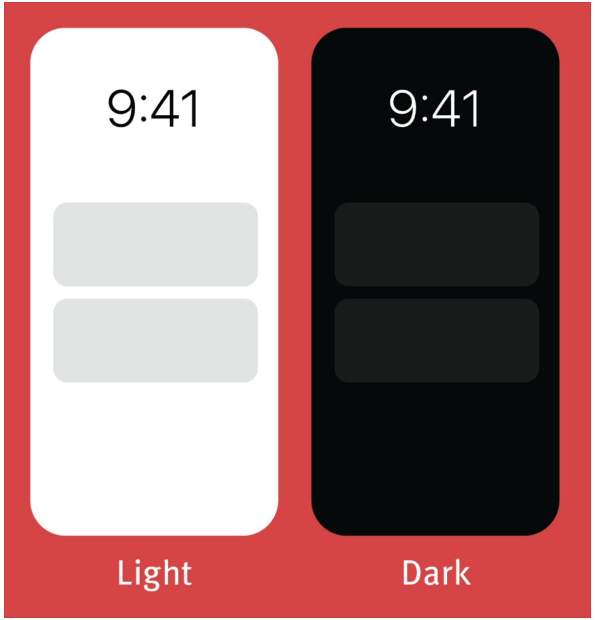


Corpus: Landing pages of 300 Alexa sites

Root Cause: **JavaScript Induced Non-Determinism**

Resources fetched different from **crawled → Poor page fidelity**


modern portfolio
About
Portfolio
Services
Blog
Contact
Welcome to my website.
My name is Brian Gardner and I am freelance web designer living in the suburbs of Chicago. I have been a Starbucks addict since 1974, and have been known to love the music sung by Sarah McLachlan. I write poetry and am a nomad theorist.
View my portfolio.
Portfolio Post #2
Portfolio Post #3



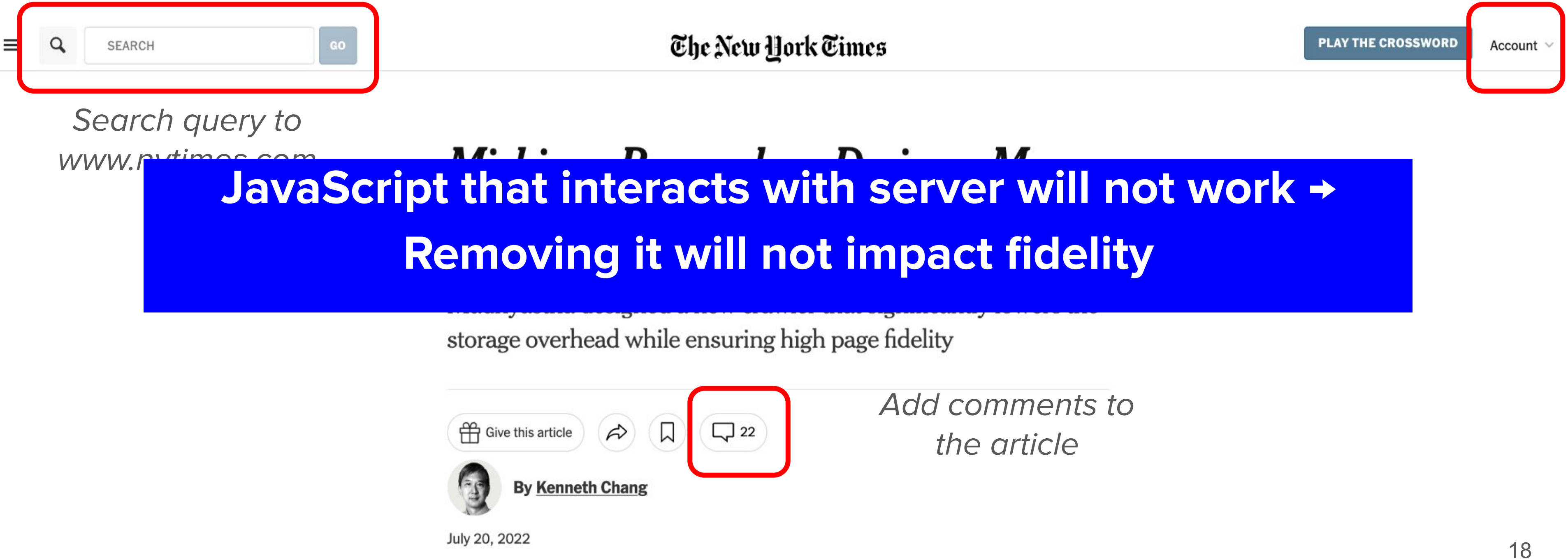
Outline

- ★ Challenges faced by web archives
- ★ Understanding the root cause
- ★ **Our insights and design of a new crawler**
- ★ Evaluation

Key Insight: Archived Page Differs from Live Page

A. No back-end origin server

Account login an
subscription



storage overhead while ensuring high page fidelity

July 20, 2022

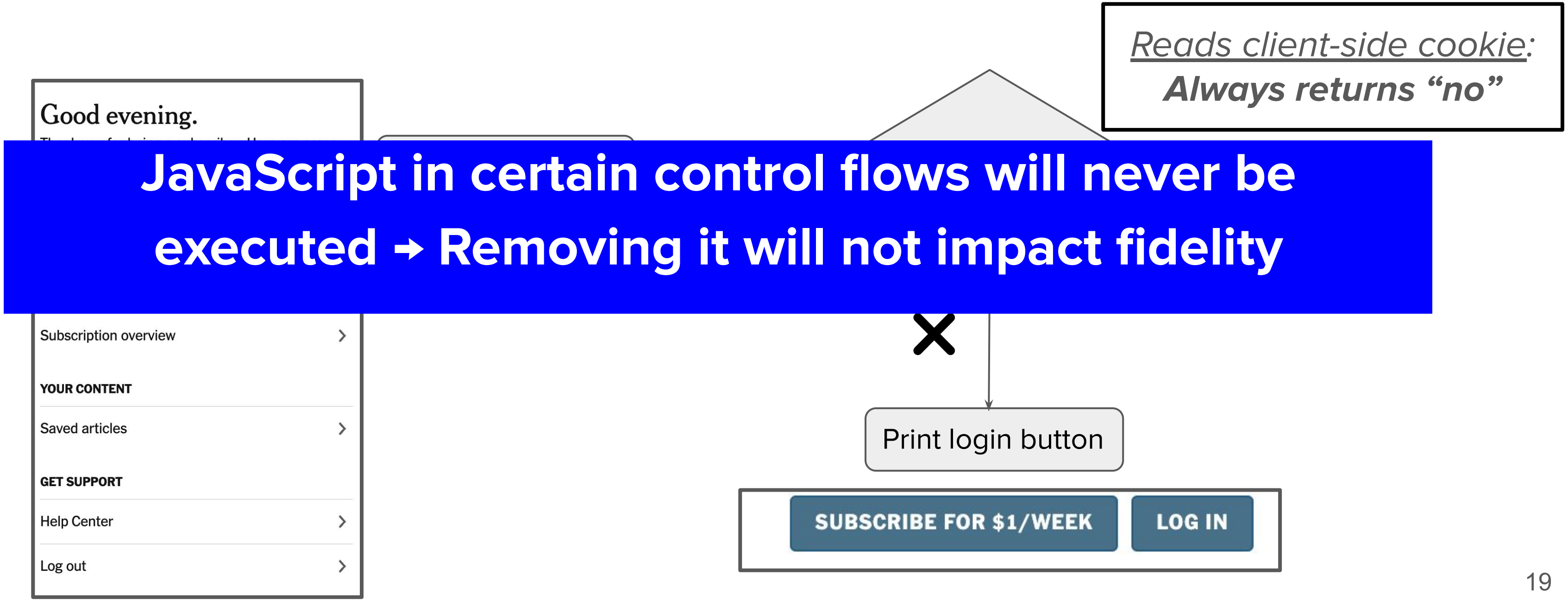
Search query to www.nytimes.com

JavaScript that interacts with server will not work → Removing it will not impact fidelity

Add comments to the article

Key Insight: Archived Page Differs from Live Page

B. Certain sources of non-determinism are absent



Good evening.

Subscription overview

YOUR CONTENT

Saved articles

GET SUPPORT

Help Center

Log out

Reads client-side cookie:
Always returns "no"

JavaScript in certain control flows will never be executed → Removing it will not impact fidelity

Jawa: A New Web Archive Crawler

Improve page fidelity

DAILY CALLER

'Peaceful Protesters'

'Wild Goose Chase': Strzok Says Steele Dossier Was 'A Problem' For The Trump-Russia Probe

President Trump Calls On Congress To Release $300 Billion In Relief Funds To Americans

Here's How Democrats And The Media Have Responded To The Months Of Civil Unrest Following George Floyd's Death

Will Democrats Accept Defeat in

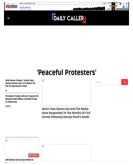
- ➔ **Eliminate impact of non-determinism**

Reduce storage cost

- ➔ Identify and remove non-functional code
- ➔ **Identify and remove unreachable code**

Jawa: A New Web Archive Crawler

Improve page fidelity



DAILY CALLER

'Peaceful Protesters'

'Wild Goose Chases': Strzok Says Steele Dossier Has 'A Problem' For The Trump-Russia Probe

President Trump Calls On Congress To Release $300 Billion In Relief Funds To Americans

Here's How Democrats And The Media Have Responded To The Months Of Civil Unrest Following George Floyd's Death

Search...

PATRIOTS

Will Democrats Accept Defeat In

➔ **Eliminate impact of non-determinism**

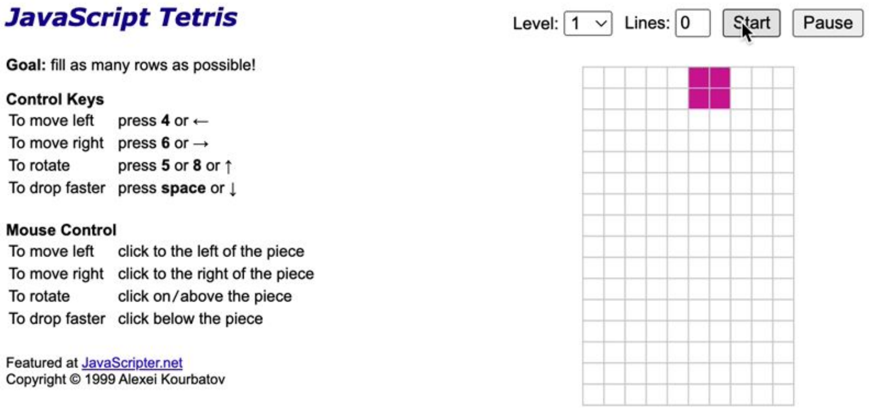
Reduce storage cost

➔ Identify and remove non-functional code

➔ **Identify and remove unreachable code**

Make JavaScript Execution **Completely** Deterministic

Original page



JavaScript Tetris

Level: 1 Lines: 0

Goal: fill as many rows as possible!

Control Keys

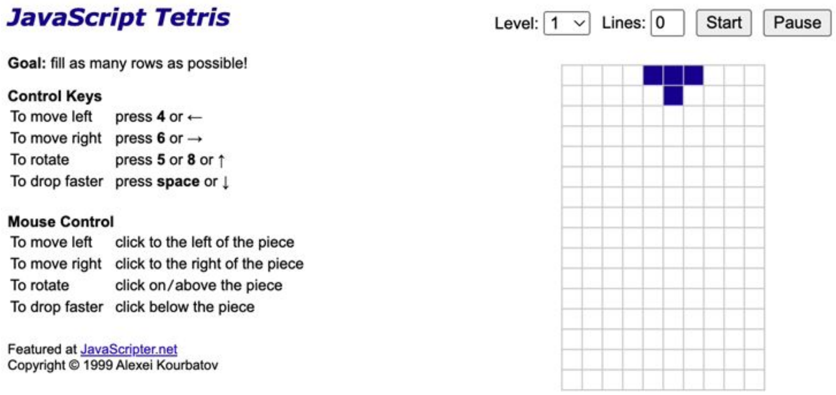
To move left	press **4** or ←
To move right	press **6** or →
To rotate	press **5** or **8** or ↑
To drop faster	press **space** or ↓

Mouse Control

To move left	click to the left of the piece
To move right	click to the right of the piece
To rotate	click on/above the piece
To drop faster	click below the piece

Featured at JavaScripter.net
Copyright © 1999 Alexei Kourbatov

Deterministic page (broken)



JavaScript Tetris

Level: 1 Lines: 0

Goal: fill as many rows as possible!

Control Keys

To move left	press **4** or ←
To move right	press **6** or →
To rotate	press **5** or **8** or ↑
To drop faster	press **space** or ↓

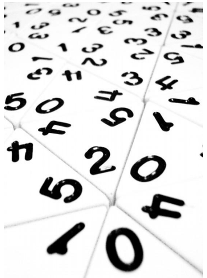
Mouse Control

To move left	click to the left of the piece
To move right	click to the right of the piece
To rotate	click on/above the piece
To drop faster	click below the piece

Featured at JavaScripter.net
Copyright © 1999 Alexei Kourbatov

Understand How Non-determinism Impacts Resources Fetched

Date, Math.random, Performance (DRP)



Client characteristics

Analyze impact on control flow on 3000 pages

No impact on control flow

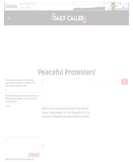
➔ retain the non-determinism

Influences control flow impacting resource fetches

➔ Eliminate non-determinism

Jawa: A New Web Archive Crawler

Improve page fidelity



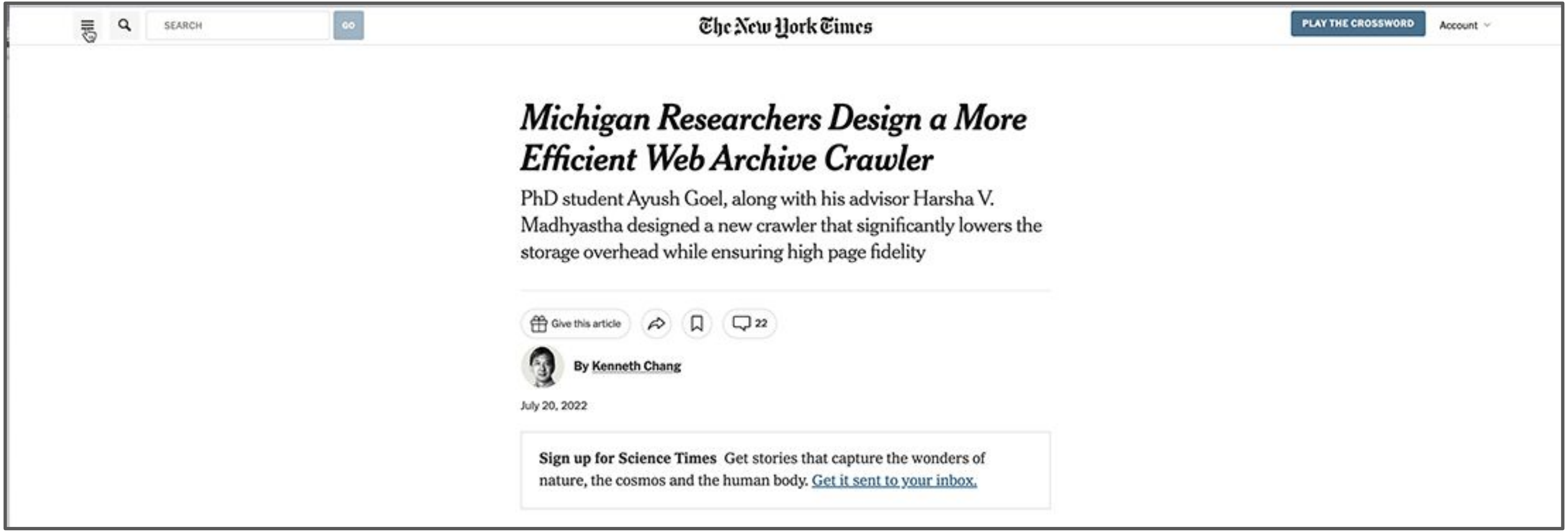
→ **Eliminate impact of non-determinism**

Reduce storage cost

→ Identify and remove non-functional code

→ **Identify and remove unreachable code**

Insufficient to Save Only JS Executed While Crawling



Michigan Researchers Design a More Efficient Web Archive Crawler

PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity

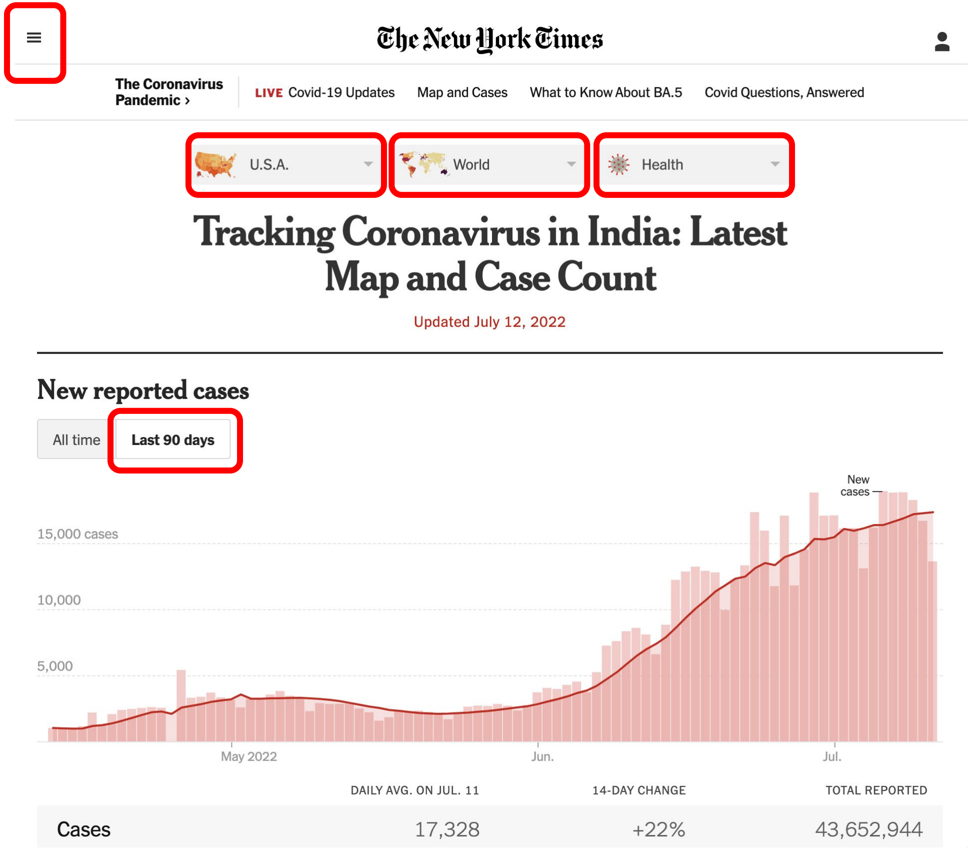
By Kenneth Chang

July 20, 2022

Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.

Challenge: Code Executed Depends on How User Interacts

Challenge: Code Executed Depends on **Order**



The New York Times

The Coronavirus Pandemic ›

LIVE Covid-19 Updates

Map and Cases

What to Know About BA.5

Covid Questions, Answered

Tracking Coronavirus in India: Latest Map and Case Count

Updated July 12, 2022

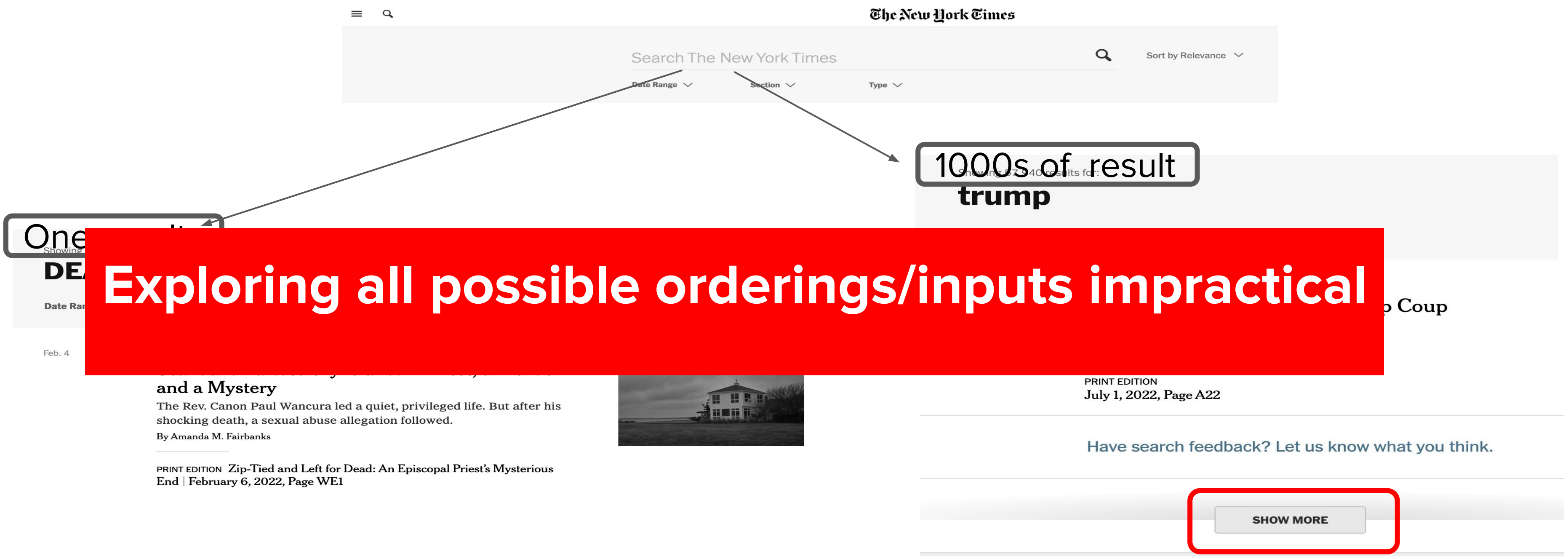
New reported cases

	DAILY AVG. ON JUL. 11	14-DAY CHANGE	TOTAL REPORTED
Cases	17,328	+22%	43,652,944

Challenge: Code Executed Depends on **Order**

- ➔ **Read-write dependencies** between different events on page
    - ◆ **JavaScript/DOM state**
- ➔ **Can not predict** order of events

Challenge: Code Executed Depends on **Input**



The New York Times

Search The New York Times

Sort by Relevance

Date Range

Section

Type

1000s of result

Showing 57,940 results for:

trump

One

Showing

DE

Date Ran

Feb. 4

and a Mystery

The Rev. Canon Paul Wancura led a quiet, privileged life. But after his shocking death, a sexual abuse allegation followed.

By Amanda M. Fairbanks

PRINT EDITION Zip-Tied and Left for Dead: An Episcopal Priest's Mysterious End | February 6, 2022, Page WE1

Coup

PRINT EDITION
July 1, 2022, Page A22

Have search feedback? Let us know what you think.

SHOW MORE

Exploring all possible orderings/inputs impractical

Analysis Framework: Understand State Dependencies

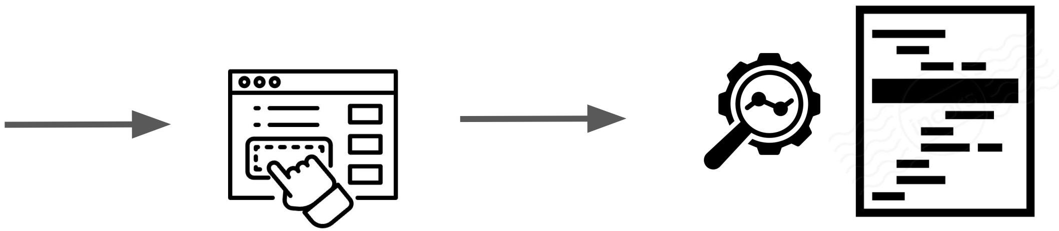
Inject instrumentation code

```
function foo(a,b){
     var p = Proxy(win,handler)
     p.c = _(a)+_(b);
     return _ret(true);
}

foo(1,2);
```

Trigger each interaction

Monitor state accesses



Corpus: 3000 pages from 300 Alexa sites

Findings From the Analysis

- ➔ Order **mostly** influences coverage for **analytics events**

Both categories of events irrelevant for archived pages

- ➔ Input **mostly** influences coverage for **text-based events**

Implication:

Estimate code coverage with **single execution of events with default inputs**

Outline

- ★ Challenges faced by web archives
- ★ Understanding the root cause
- ★ Our insights and design of a new crawler
- ★ **Evaluation**

Evaluation

Storage

- Storage for web resources
- Storage for crawling/serving indices

Fidelity

- Failed Resource fetches
- Visual comparison
- Functional interactions

Throughput

- Crawling throughput
- Overhead of each technique

Evaluation

Storage

- **Storage for web resources**
- Storage for crawling/serving indices

Fidelity

- **Failed Resource fetches**
- Visual comparison
- Functional interactions

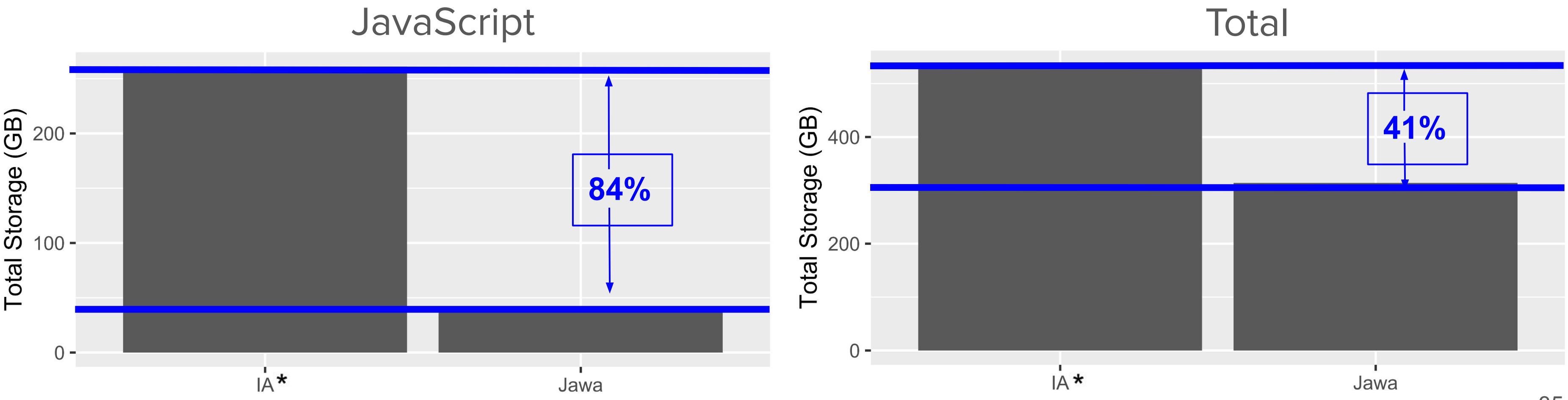
Throughput

- Crawling throughput
- Overhead of each technique

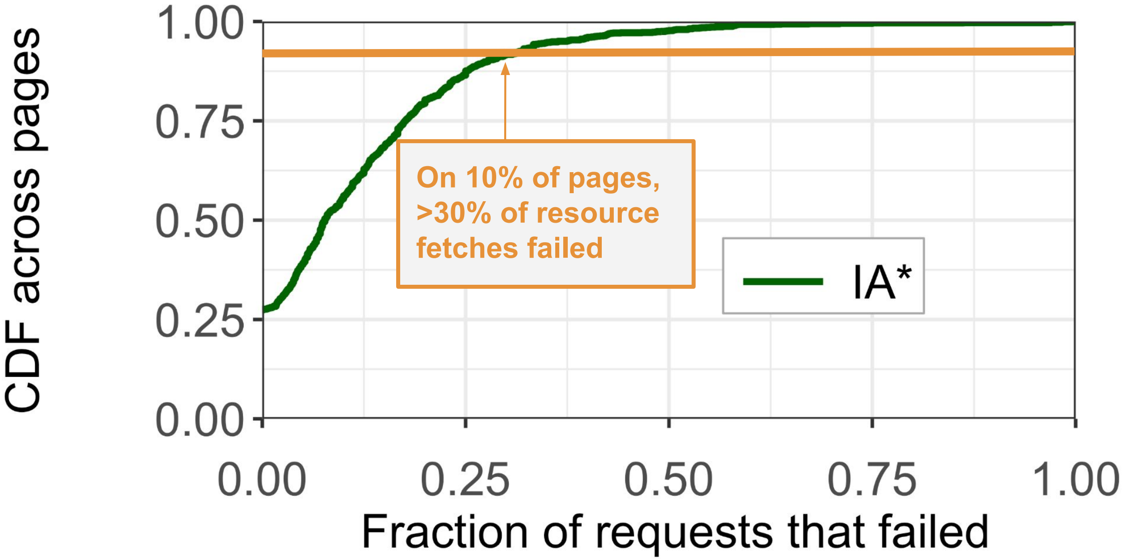
Storage: Jawa Reduces Storage Overhead by **41%**

Corpus:

<u>1 million snapshots</u> of pages on 300 sites archived by Internet Archive



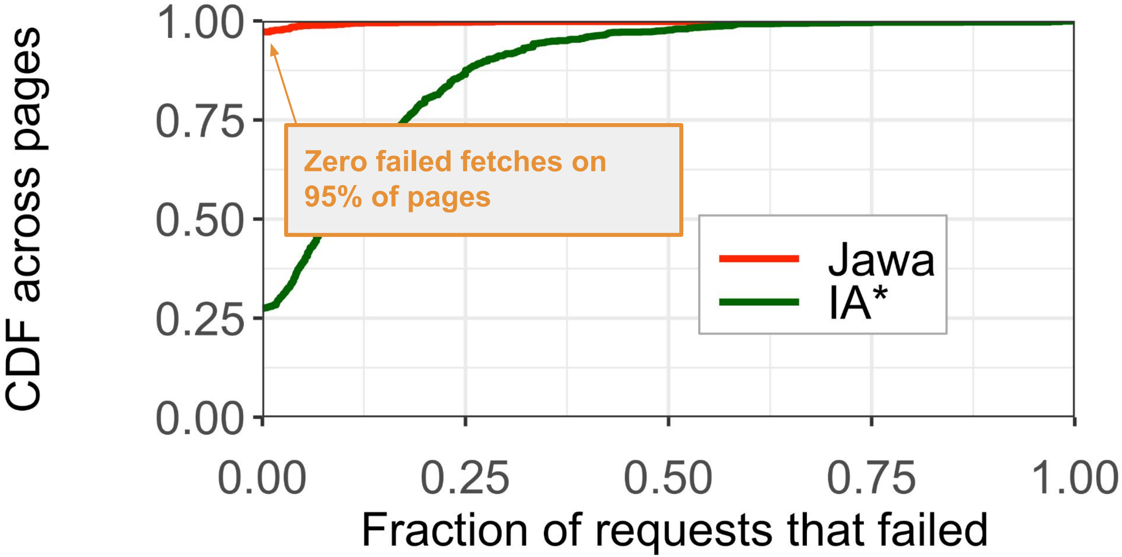
Fidelity: Jawa Eliminates Almost All Failed Resource Fetches



On 10% of pages, >30% of resource fetches failed

Corpus: 3000 web pages from 300 sites

Fidelity: Jawa Eliminates Almost All Failed Resource Fetches



Zero failed fetches on 95% of pages

Corpus: 3000 web pages from 300 sites

Conclusion

1. JavaScript on web pages negatively impacts web archival

2. Fundamental differences b/w live and archived pages can be exploited to overcome such negative impacts

https://github.com/goelayu/Jawa

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

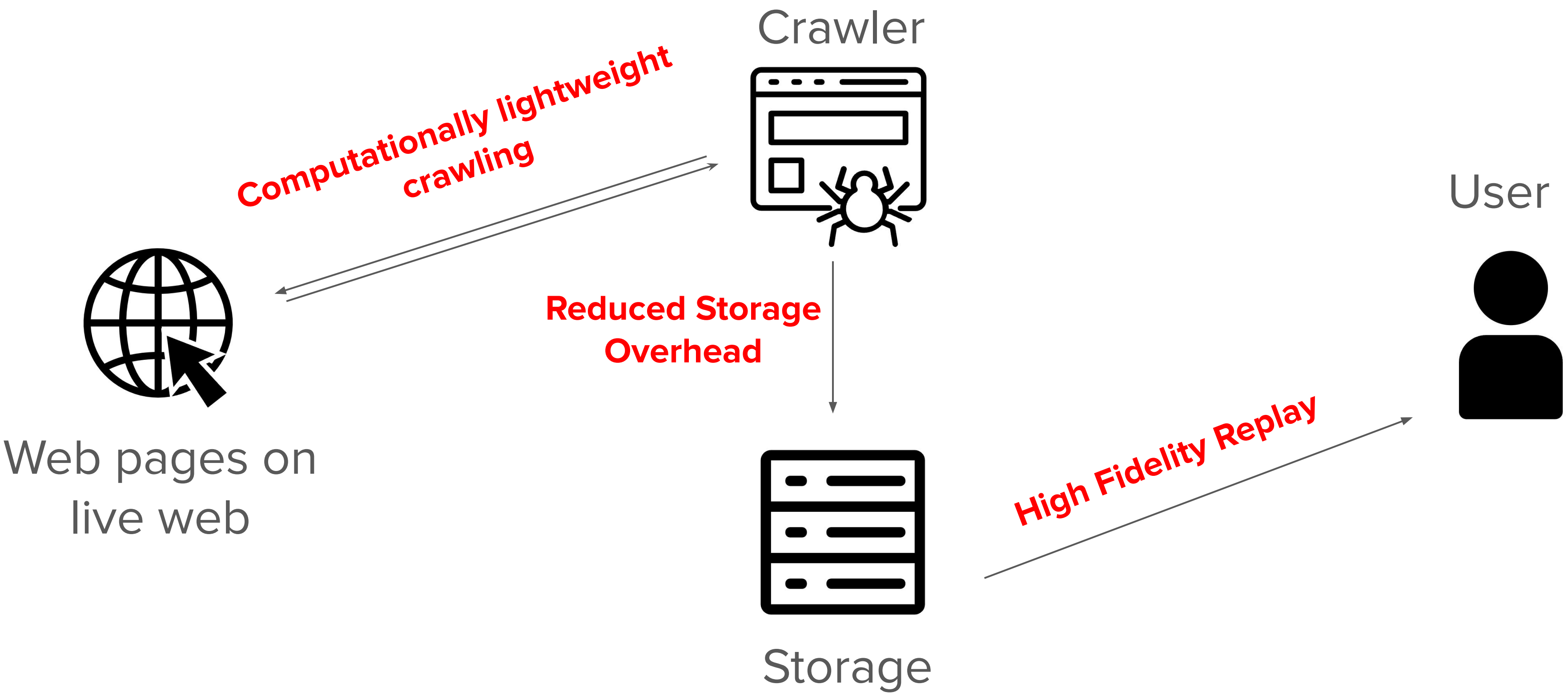
ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

goelayu@umich.edu

Evaluation.

Backup

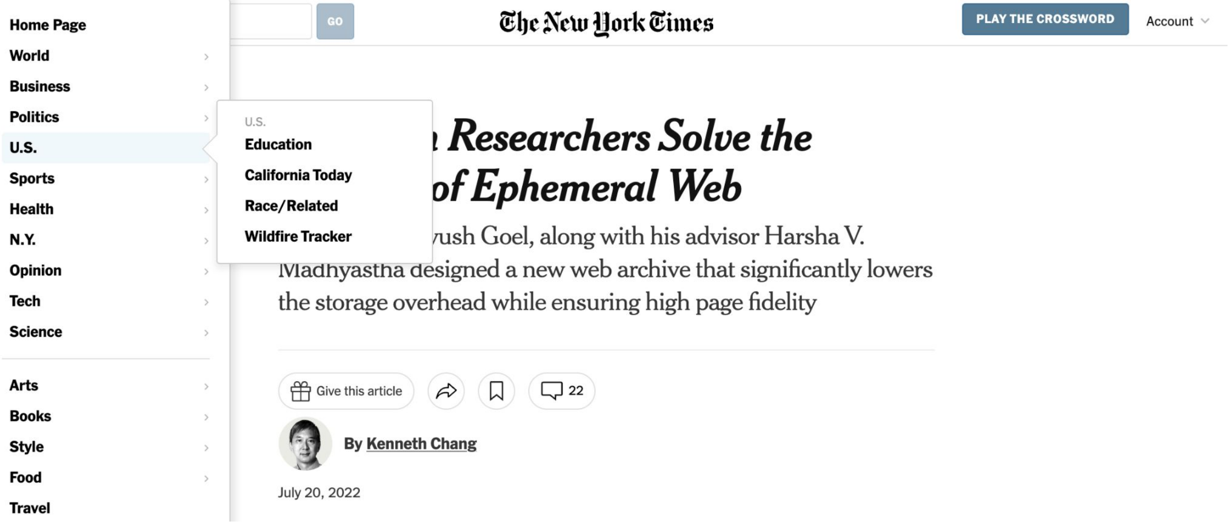
Jawa: A New Web Archival Crawler



Improve Page Fidelity: Strawman #1

Screen capture

✗ No client interactions



Home Page
World
Business
Politics
U.S.
Sports
Health
N.Y.
Opinion
Tech
Science
Arts
Books
Style
Food
Travel

U.S.
Education
California Today
Race/Related
Wildfire Tracker

ı Researchers Solve the
of Ephemeral Web

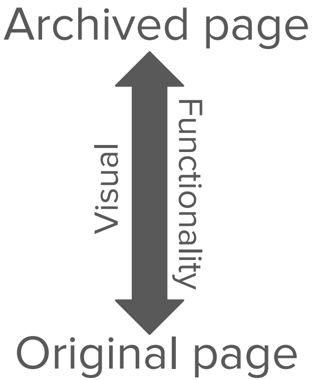
ush Goel, along with his advisor Harsha V. Madhyastha designed a new web archive that significantly lowers the storage overhead while ensuring high page fidelity

By Kenneth Chang

July 20, 2022

Jawa: A New Web Archival Crawler

High Fidelity



Archived page

Visual

Functionality

Original page

Low Cost

Reduced storage

Computationally lightweight

Web Archives to the Rescue

~700 Billion pages, ~100 Petabytes

LIBRARY

LIBRARY OF CONGRESS

UKWA

UK WEB ARCHIVE

ARQUIVO.PT

Web Archives to the Rescue

To tackle this problem of ephemeral web, various web archives have been established

LIBRARY
LIBRARY OF CONGRESS
ARCHIVE-IT
UKWA
UK WEB ARCHIVE

Web Archives to the Rescue

LIBRARY

LIBRARY OF CONGRESS

UKWA

UK WEB ARCHIVE

ARQUIVO.PT

Web Archives to the Rescue

LIBRARY
LIBRARY
OF CONGRESS

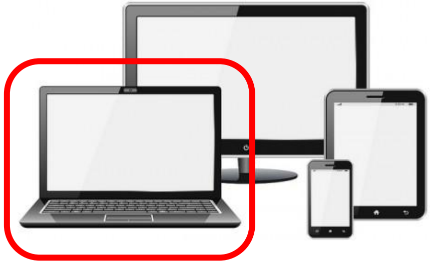
UKWA
UK WEB ARCHIVE

ARQUIVO.PT

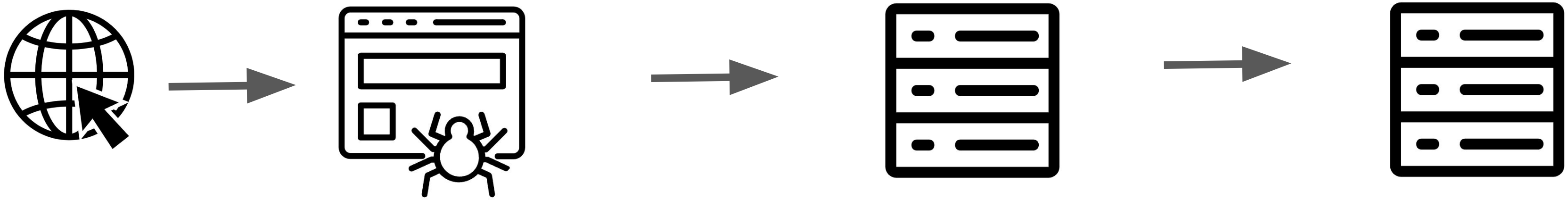
Web Archives to the Rescue

Approach to Fix Fidelity

- ➔ Do not make DRP APIs deterministic
- ➔ **Leverage server-side matching** techniques to account for such URL differences
- ➔ Enforce same client characteristics for loads as the ones used for crawling



How Modern Web Archives Operate



Crawling

- Initialize w/ list of URLs
- Visit & download all web resources

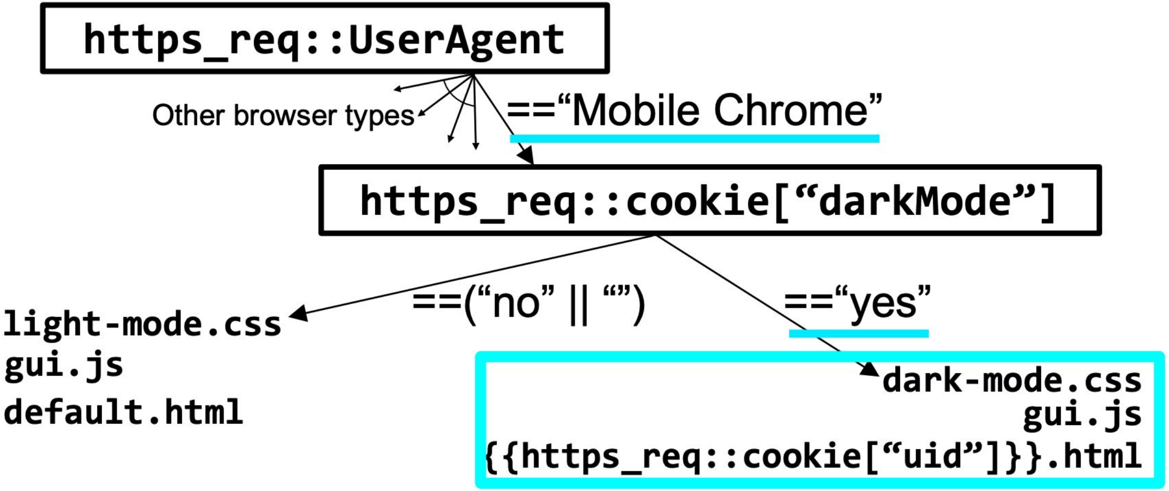
Storing

- For each resource store headers, payload.
- Apply deduplication

Replay

- Export timestamps of snapshots
- Replay each page by iteratively fetching resources

Non-determinism Resulting in Poor Fidelity



https_req::UserAgent

Other browser types

==“Mobile Chrome”

https_req::cookie[“darkMode”]

==(“no” || “”)

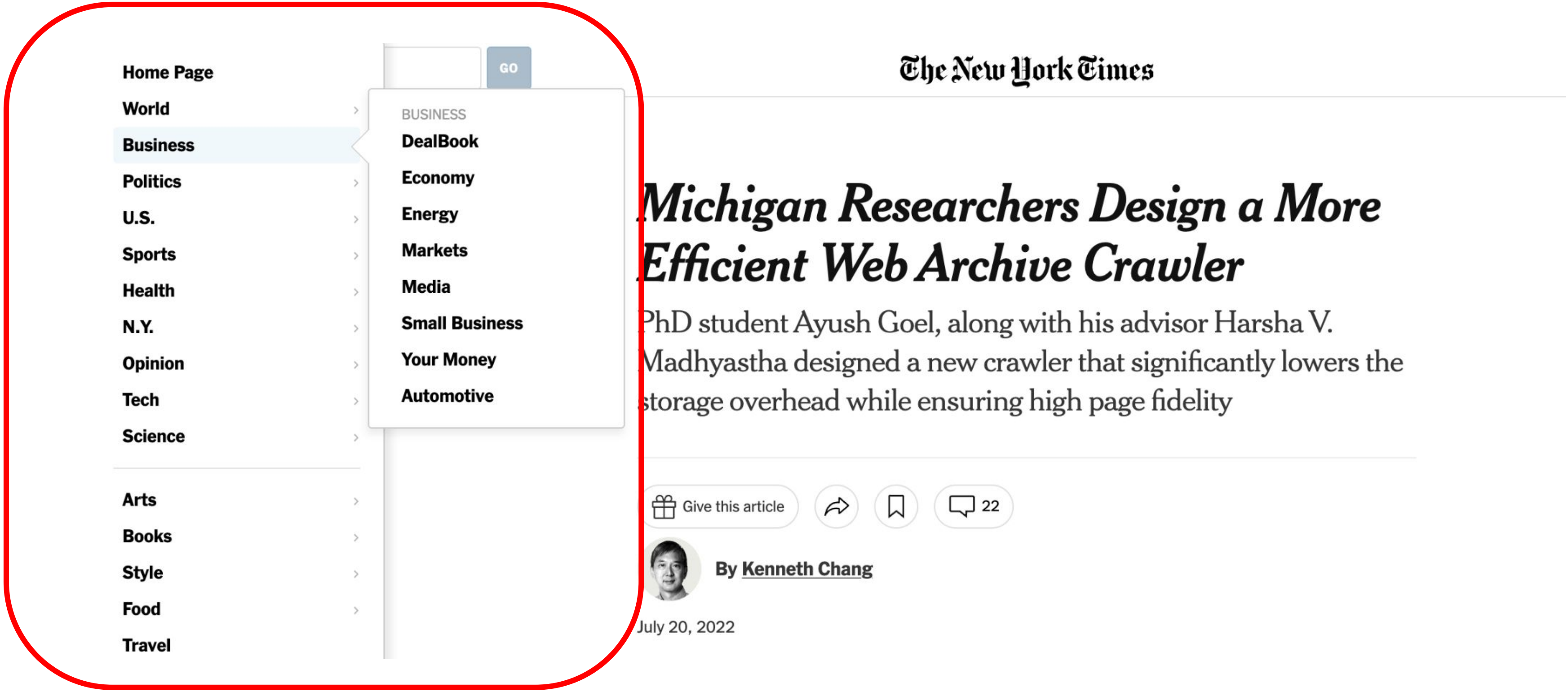
light-mode.css
gui.js
default.html

==“yes”

dark-mode.css
gui.js
{{https_req::cookie[“uid”]}}.html

Taken from Ronny et al., NSDI’21

Strawman: Preserve Code Executed **During Page Load**



Michigan Researchers Design a More Efficient Web Archive Crawler

PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity

By Kenneth Chang

July 20, 2022

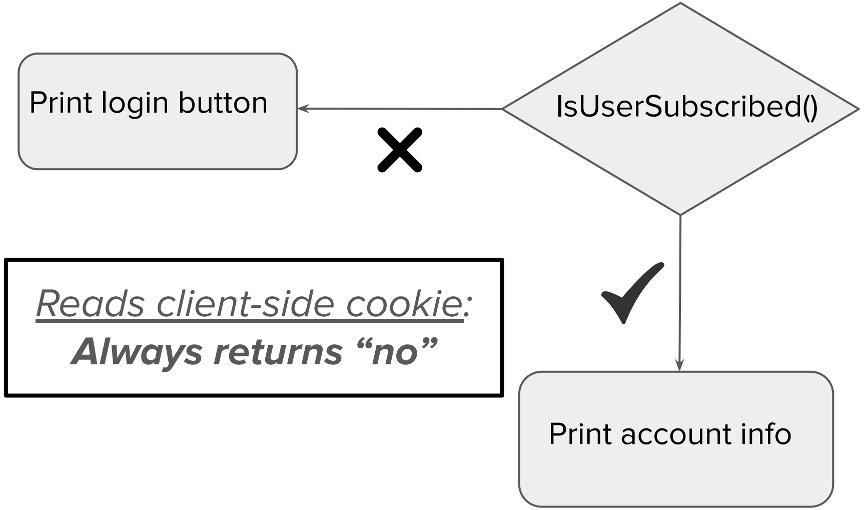
Make JavaScript Execution **Completely** Deterministic

Original page

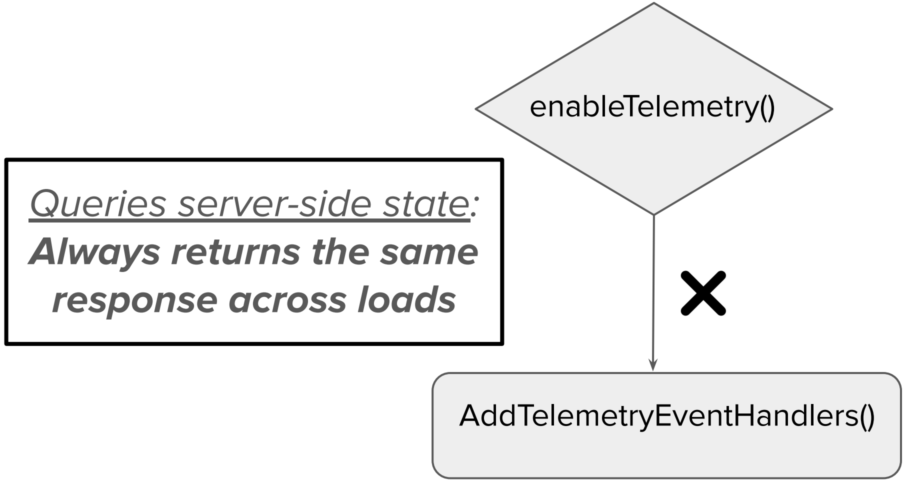
Deterministic page (broken)

Key Insight #1: Archived Page ≠ Live Page

B. Certain sources of non-determinism are absent



Reads client-side cookie:
Always returns "no"



Queries server-side state:
Always returns the same response across loads

≠