Detecting and Diagnosing Errors in Serving Archived Web Pages

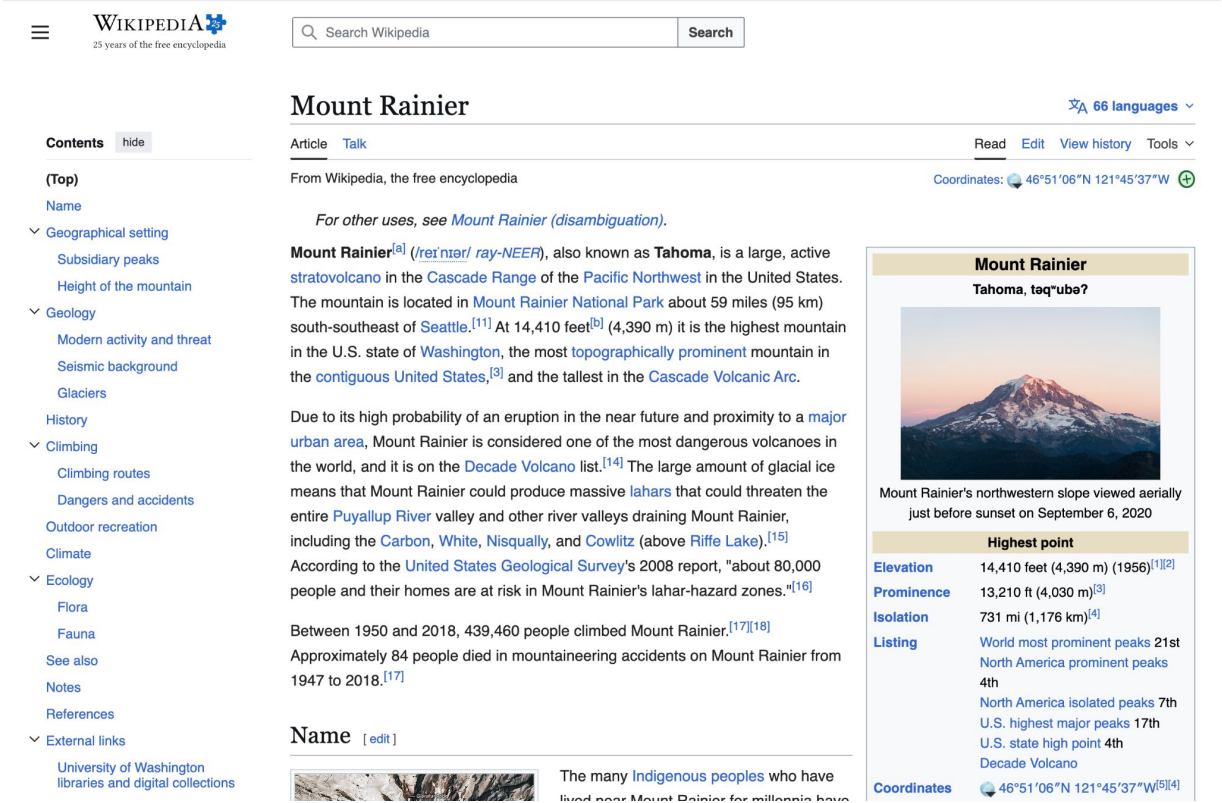
Jingyuan Zhu
University of Michigan

Huanchen Sun
University of Southern California

Harsha V. Madhyastha
University of Southern California

NSDI'26
Renton, WA, USA

Planning my trip to Seattle …



WIKIPEDIA
25 years of the free encyclopedia

Search Wikipedia | Search

Mount Rainier

66 languages

Article | Talk | Read | Edit | View history | Tools

From Wikipedia, the free encyclopedia

Coordinates: 46°51′06″N 121°45′37″W

Contents hide

- (Top)
- Name
- Geographical setting
  - Subsidiary peaks
  - Height of the mountain
- Geology
  - Modern activity and threat
  - Seismic background
  - Glaciers
- History
- Climbing
  - Climbing routes
  - Dangers and accidents
- Outdoor recreation
- Climate
- Ecology
  - Flora
  - Fauna
- See also
- Notes
- References
- External links
  - University of Washington libraries and digital collections

For other uses, see Mount Rainier (disambiguation).

Mount Rainier[a] (/reɪˈnɪər/ *ray-NEER*), also known as **Tahoma**, is a large, active stratovolcano in the Cascade Range of the Pacific Northwest in the United States. The mountain is located in Mount Rainier National Park about 59 miles (95 km) south-southeast of Seattle.[11] At 14,410 feet[b] (4,390 m) it is the highest mountain in the U.S. state of Washington, the most topographically prominent mountain in the contiguous United States,[3] and the tallest in the Cascade Volcanic Arc.

Due to its high probability of an eruption in the near future and proximity to a major urban area, Mount Rainier is considered one of the most dangerous volcanoes in the world, and it is on the Decade Volcano list.[14] The large amount of glacial ice means that Mount Rainier could produce massive lahars that could threaten the entire Puyallup River valley and other river valleys draining Mount Rainier, including the Carbon, White, Nisqually, and Cowlitz (above Riffe Lake).[15] According to the United States Geological Survey's 2008 report, "about 80,000 people and their homes are at risk in Mount Rainier's lahar-hazard zones."[16]

Between 1950 and 2018, 439,460 people climbed Mount Rainier.[17][18] Approximately 84 people died in mountaineering accidents on Mount Rainier from 1947 to 2018.[17]

Name [edit]

The many Indigenous peoples who have

Mount Rainier

Tahoma, təqʷubəʔ

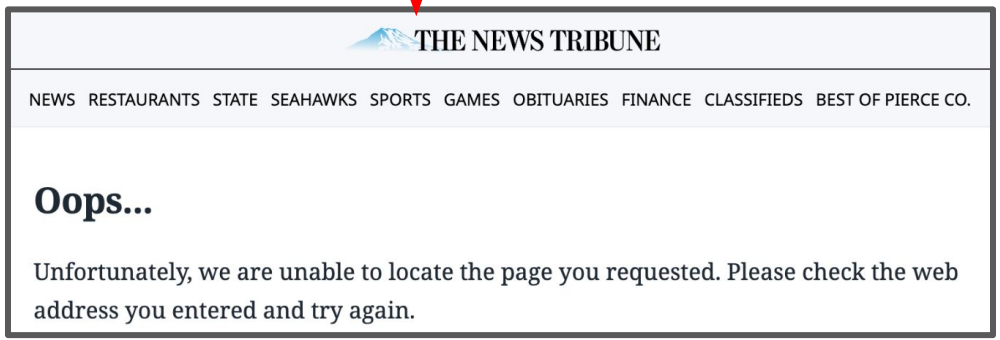
Mount Rainier's northwestern slope viewed aerially just before sunset on September 6, 2020

Highest point	
Elevation	14,410 feet (4,390 m) (1956)[1][2]
Prominence	13,210 ft (4,030 m)[3]
Isolation	731 mi (1,176 km)[4]
Listing	World most prominent peaks 21st North America prominent peaks 4th North America isolated peaks 7th U.S. highest major peaks 17th U.S. state high point 4th Decade Volcano
Coordinates	46°51′06″N 121°45′37″W[5][4]

Planning my trip to Seattle …

Geothermal heat from the volcano keeps areas of both crater rims free of snow and ice, and has formed the world's largest volcanic glacier cave network within the ice-filled craters,[37] with nearly 2 mi (3.2 km) of passages.[38] A small crater

38. ^ Sandi Doughton (October 25, 2007). "Exploring Rainier's summit steam caves". *The News Tribune*. Archived from the original on September 5, 2012. Retrieved October 3, 2010.



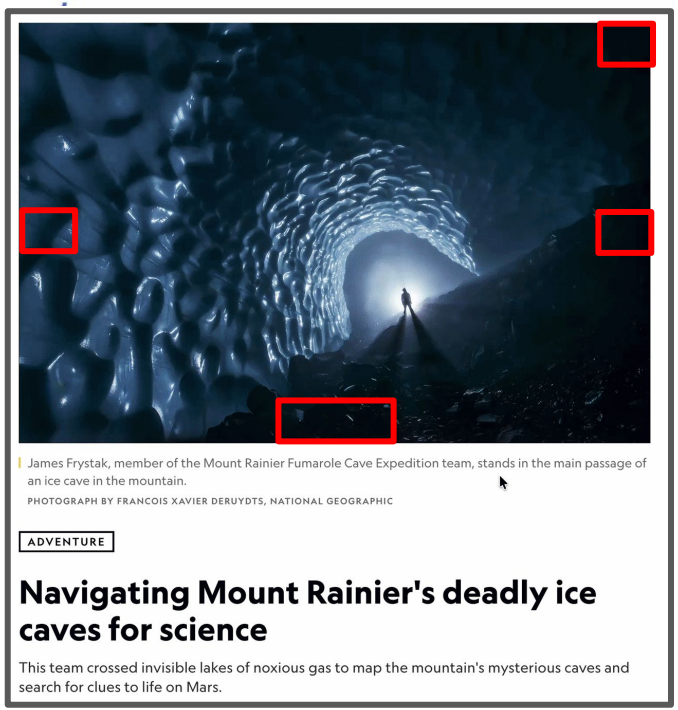
Planning my trip to Seattle …

Geothermal heat from the volcano keeps areas of both crater rims free of snow and ice, and has formed the world's largest volcanic glacier cave network within the ice-filled craters,[37] with nearly 2 mi (3.2 km) of passages.[38] A small crater

38. ^ Sandi Doughton (October 25, 2007). "Exploring Rainier's summit steam caves" *The News Tribune*. Archived from the original on September 5, 2012. Retrieved October 3, 2010.

Archived copies are not always good

3. ^ a b c Bisharat, Andrew (October 3, 2018). "Navigating Mount Rainier's deadly ice caves for science". *National Geographic*. Archived from the original on February 5, 2024. Retrieved October 16, 2024.



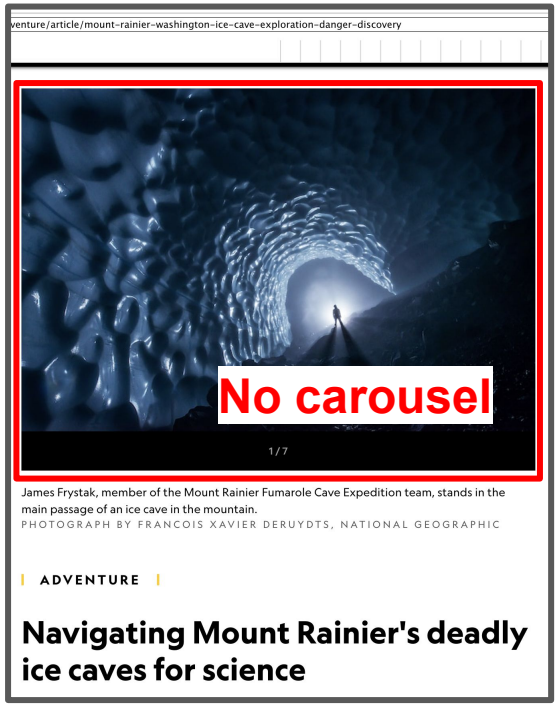
James Frystak, member of the Mount Rainier Fumarole Cave Expedition team, stands in the main passage of an ice cave in the mountain.

PHOTOGRAPH BY FRANCOIS XAVIER DERUYDTS, NATIONAL GEOGRAPHIC

ADVENTURE

Navigating Mount Rainier's deadly ice caves for science

This team crossed invisible lakes of noxious gas to map the mountain's mysterious caves and search for clues to life on Mars.



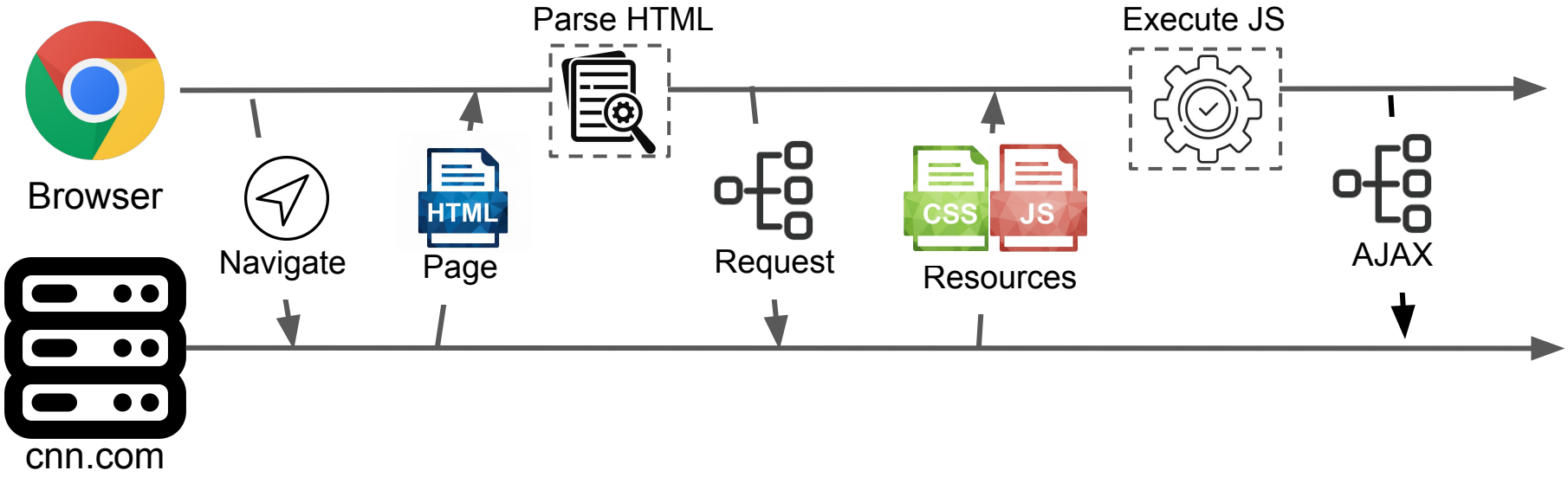
James Frystak, member of the Mount Rainier Fumarole Cave Expedition team, stands in the main passage of an ice cave in the mountain.

PHOTOGRAPH BY FRANCOIS XAVIER DERUYDTS, NATIONAL GEOGRAPHIC

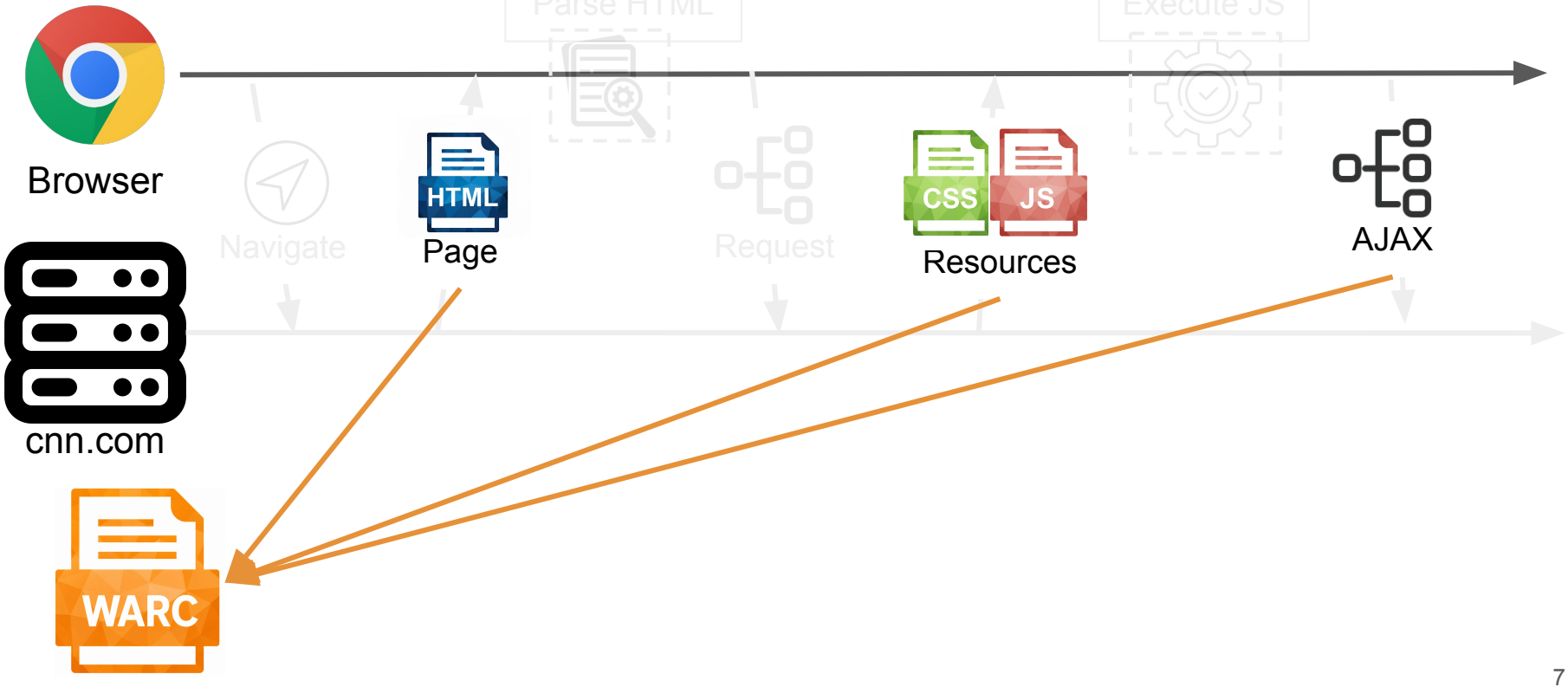
ADVENTURE

Navigating Mount Rainier's deadly ice caves for science

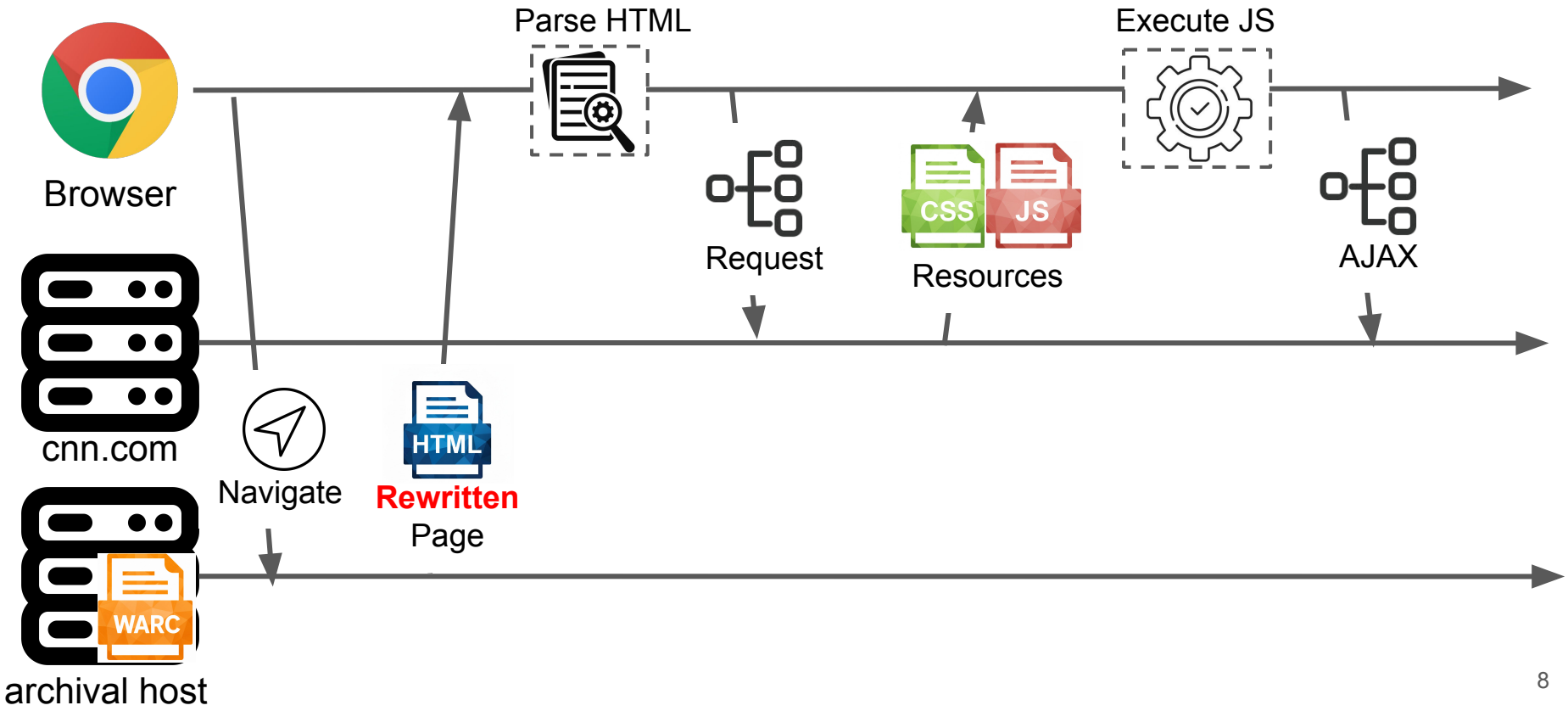
Web archives capture and serve snapshots of webpages



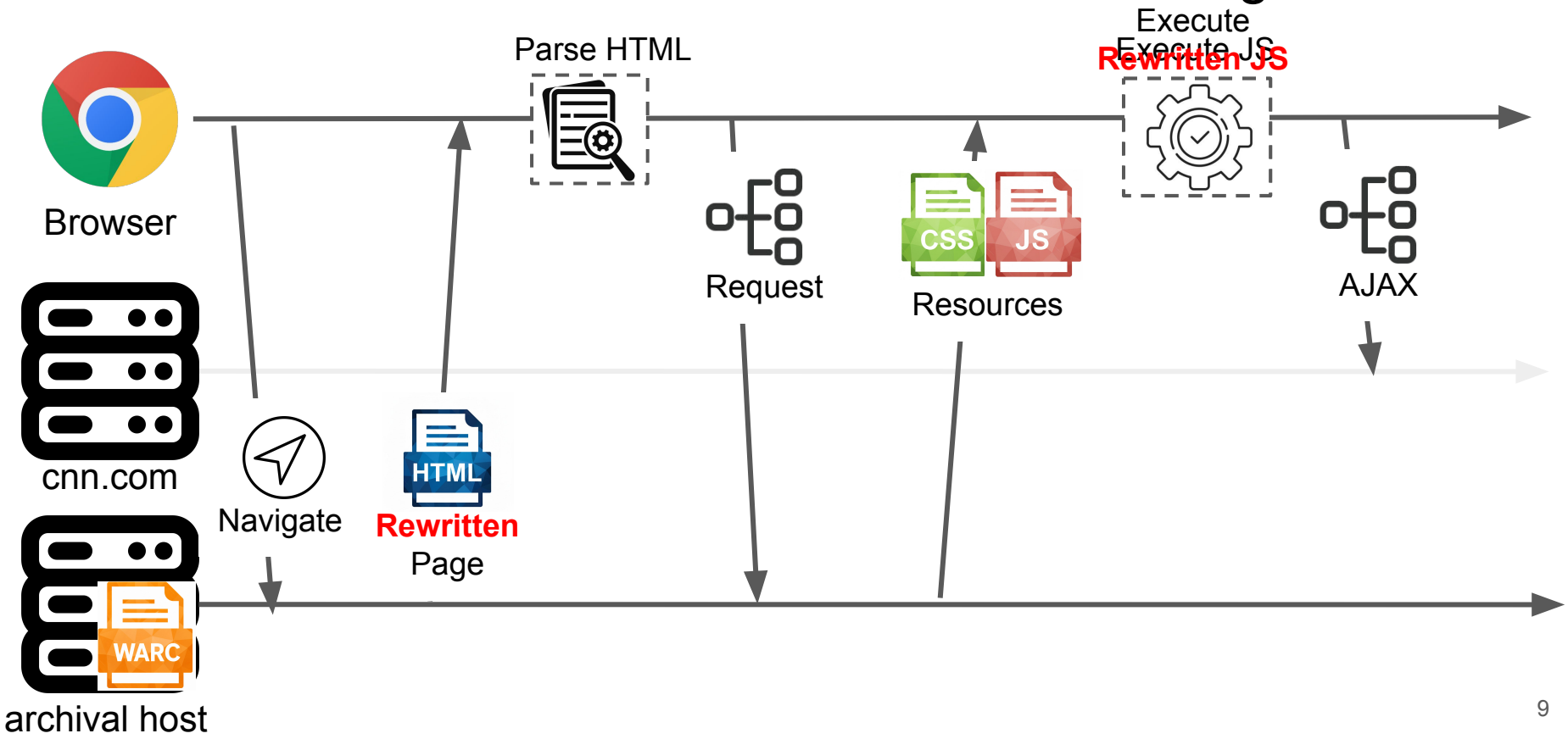
Web archives capture and serve snapshots of webpages



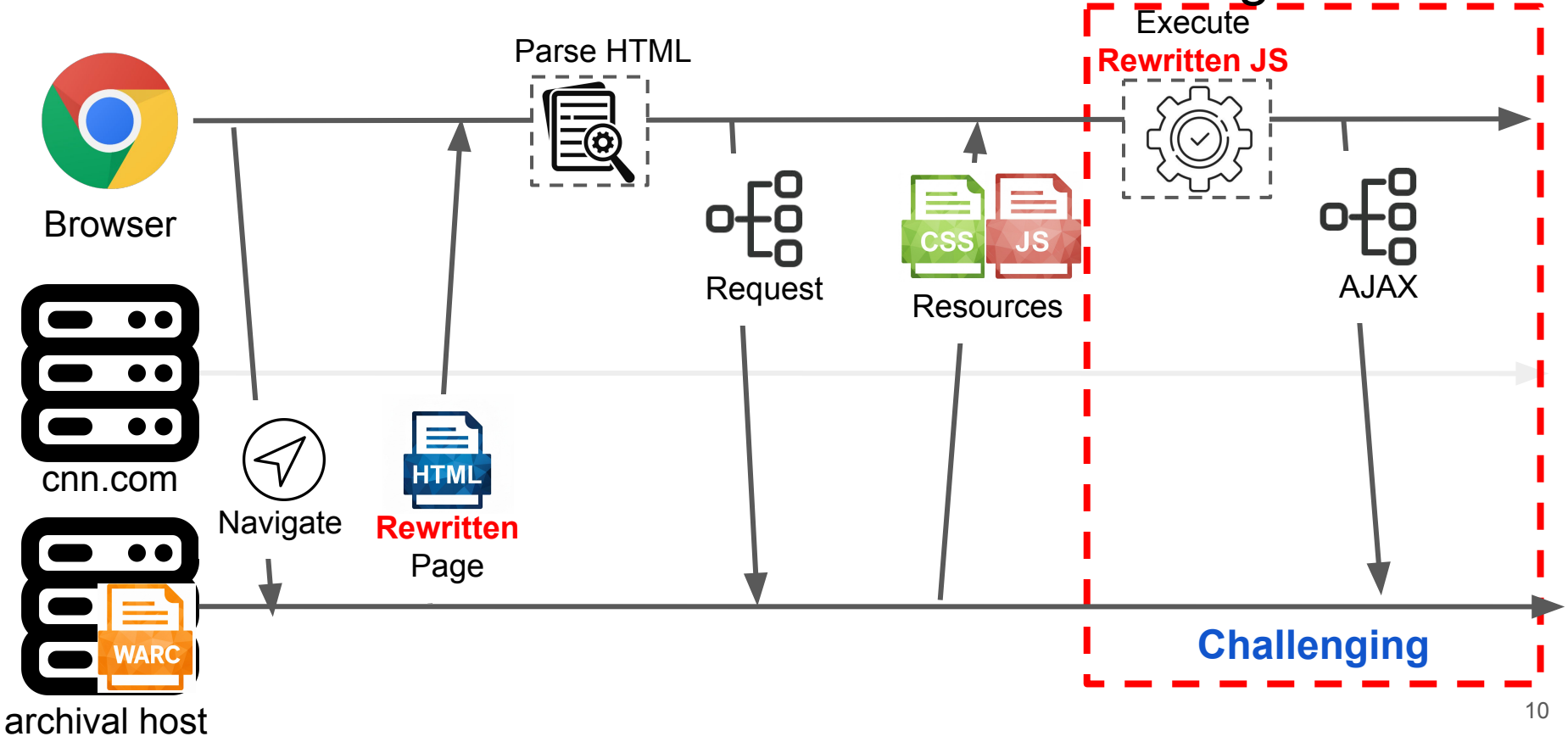
Need to rewrite crawled resources before serving them



Need to rewrite crawled resources before serving them

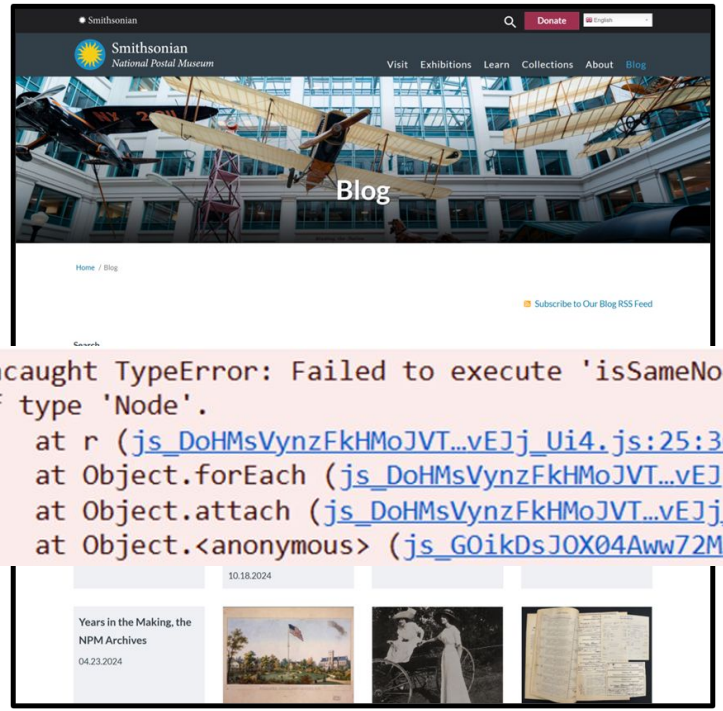


Need to rewrite crawled resources before serving them

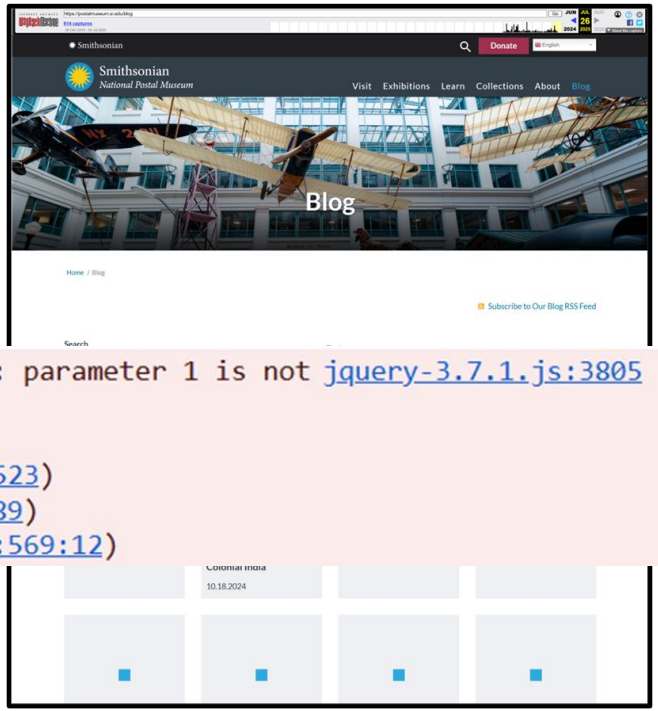


Incorrect JS rewrites lead to missing page content

Live web page



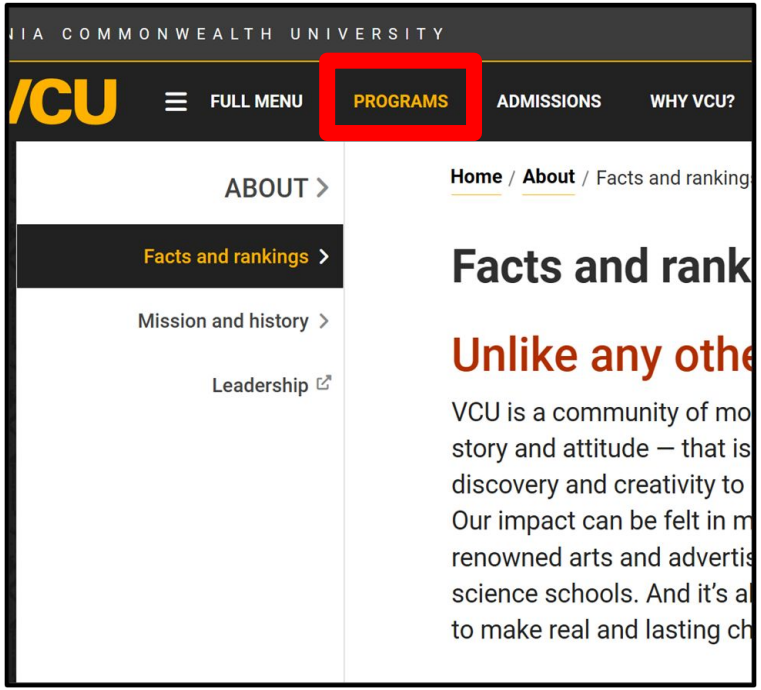
Archived copy




```
Uncaught TypeError: Failed to execute 'isSameNode' on 'Node': parameter 1 is not jquery-3.7.1.js:3805
of type 'Node'.
    at r (js_DoHMsVynzFkHMoJVT…vEJj_Ui4.js:25:3073)
    at Object.forEach (js_DoHMsVynzFkHMoJVT…vEJj_Ui4.js:25:2523)
    at Object.attach (js_DoHMsVynzFkHMoJVT…vEJj_Ui4.js:25:3489)
    at Object.<anonymous> (js_GOikDsJOX04Aww72M…RdkL0XlDo.js:569:12)
```

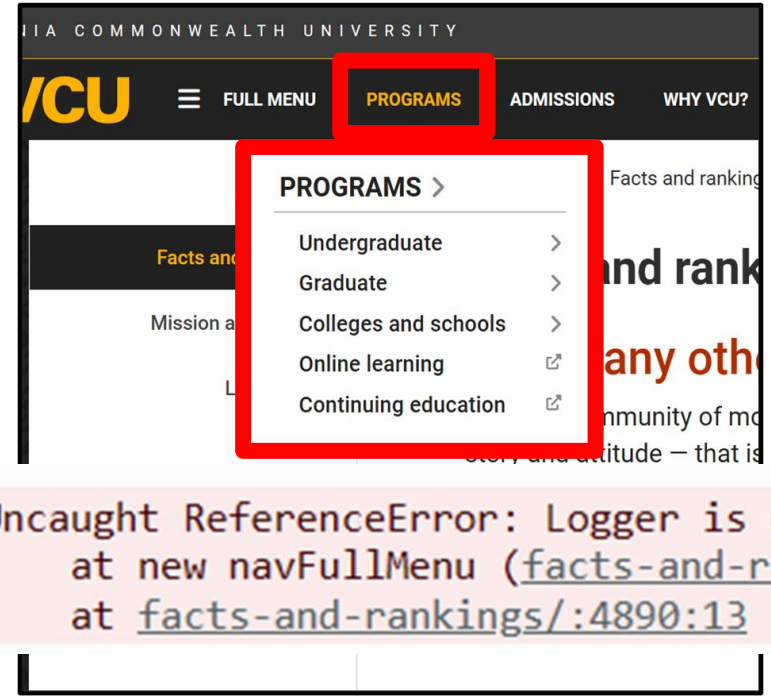
Incorrect JS rewrites break page functionality

Live web page

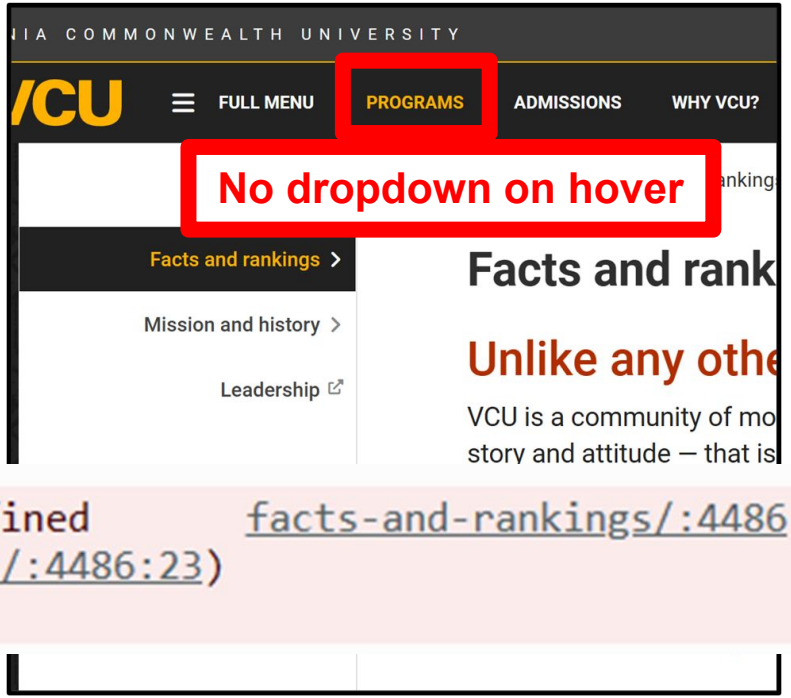


Incorrect JS rewrites break page functionality

Live web page

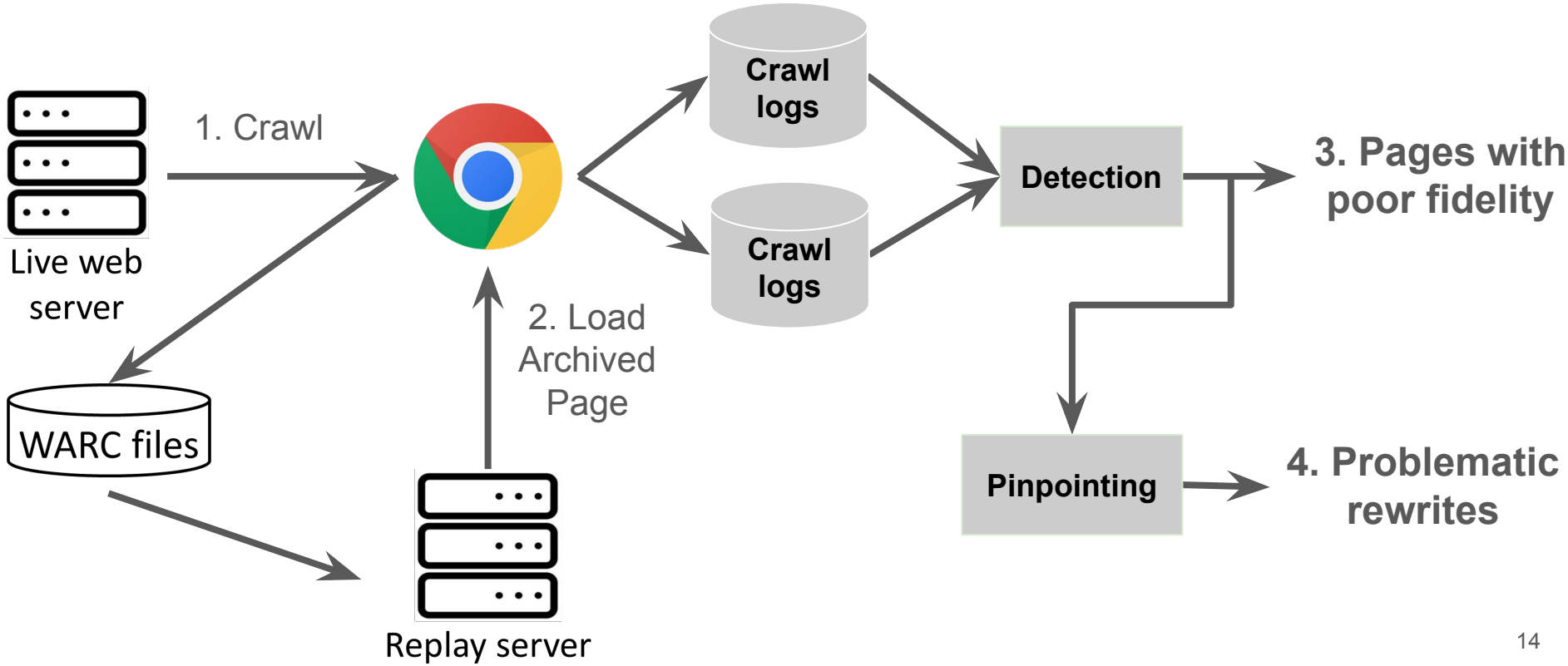


Archived copy

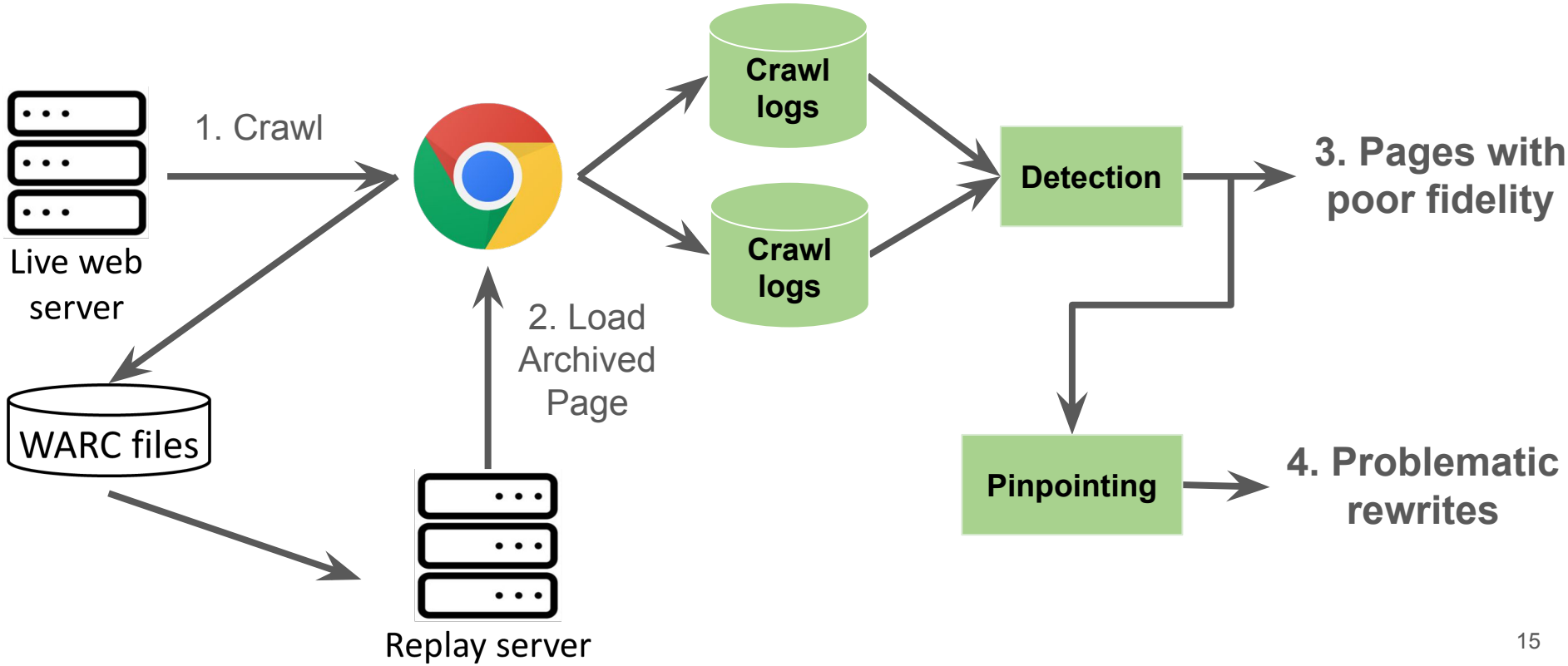



```
Uncaught ReferenceError: Logger is not defined        facts-and-rankings/:4486
    at new navFullMenu (facts-and-rankings/:4486:23)
    at facts-and-rankings/:4890:13
```

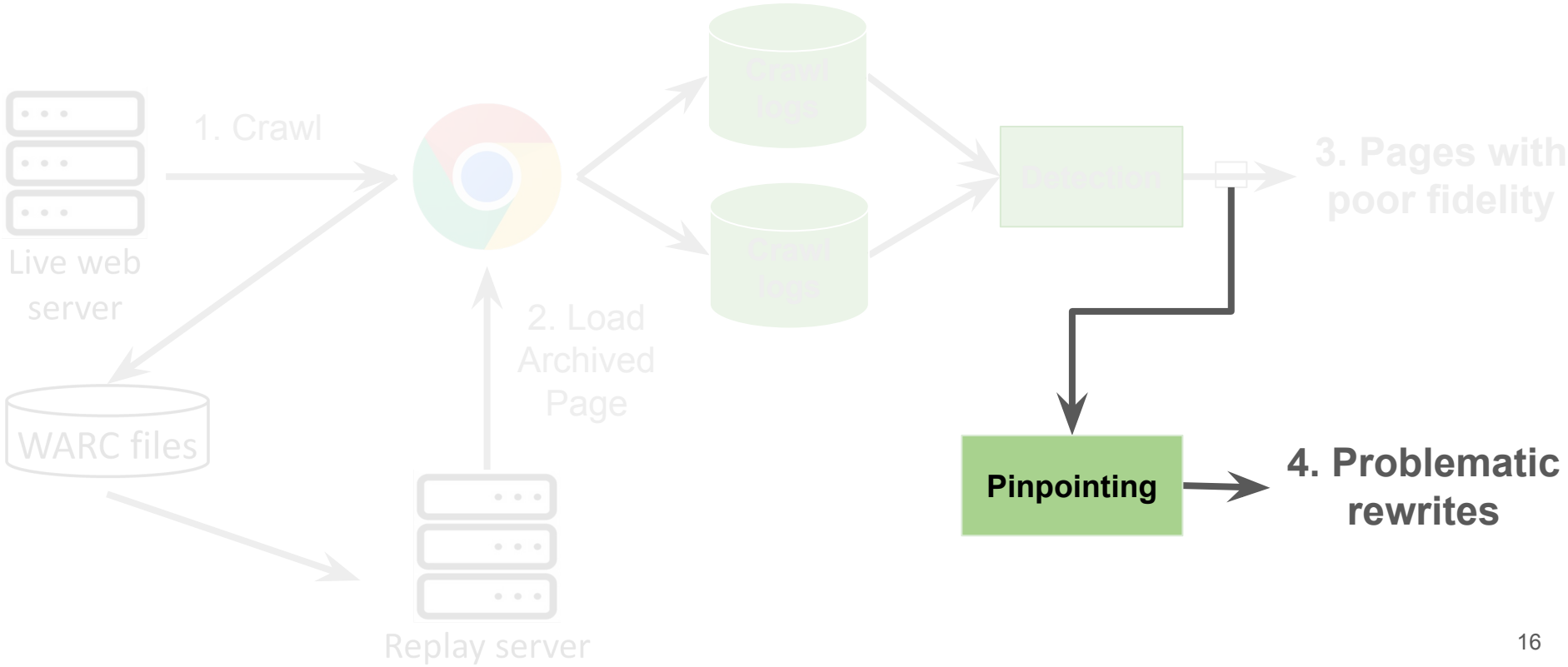
Fidelity violation detection and diagnosis



FidEx (**Fid**elity violation **Ex**poser)



FidEx (**Fid**elity violation **Ex**poser)



Runtime errors ⇏ Fidelity violations

Syntax errors for unused libraries

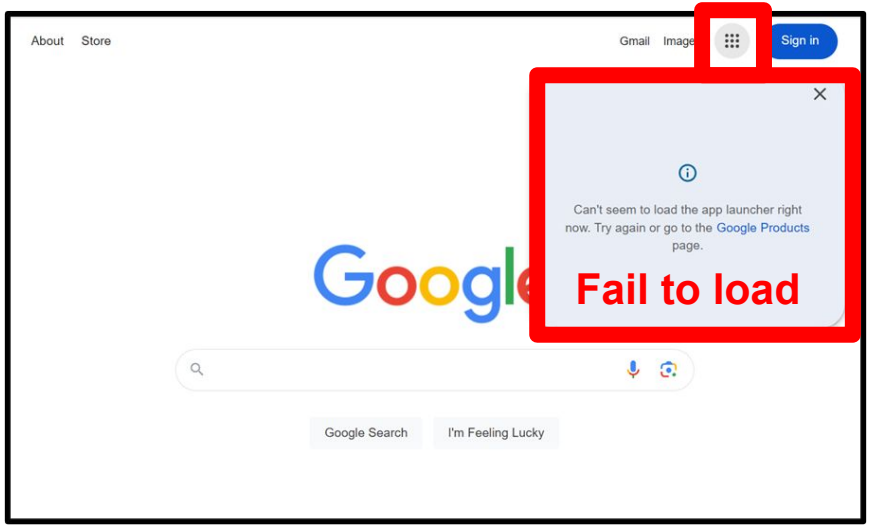
```
Uncaught SyntaxError: Unexpected token '<' (at lightbox-plus-jquery.min.js:1
lightbox-plus-jquery.min.js:1:1)
```

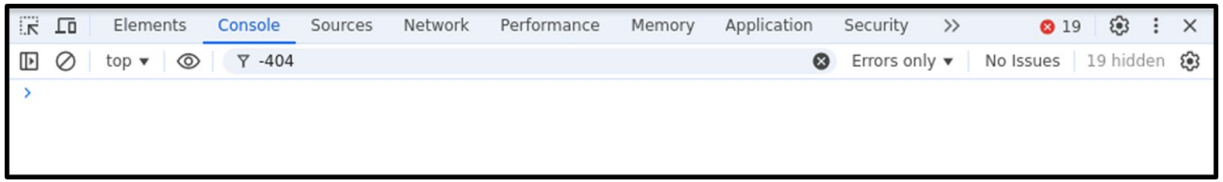
Thrown exceptions on analytical scripts

```
Uncaught TypeError: Illegal invocation                gdprscript.js?buildTime=1732220787:20
    at HTMLDocument.get (gdprscript.js?buildT…=1732220787:20:5065)
    at Wombat.defaultProxyGet (wombat.js:1351:23)
    at Object.get (wombat.js:5046:27)
    at f.getCookie (otSDKStub.js:15:13627)
    at f.readCookieParam (otSDKStub.js:15:13264)
    at f.validateIABGDPRApplied (otSDKStub.js:15:12918)
    at f.initializeIABData (otSDKStub.js:15:12517)
    at f.ensureHtmlGroupDataInitialised (otSDKStub.js:15:11891)
    at f.initConsentSDK (otSDKStub.js:15:7001)
    at new f (otSDKStub.js:15:22528)
get                              @ gdprscript.js?buildTime=1732220787:20
Wombat.defaultProxyGet           @ wombat.js:1351
get                              @ wombat.js:5046
f.getCookie                      @ otSDKStub.js:15
f.readCookieParam                @ otSDKStub.js:15
f.validateIABGDPRApplied         @ otSDKStub.js:15
f.initializeIABData              @ otSDKStub.js:15
```

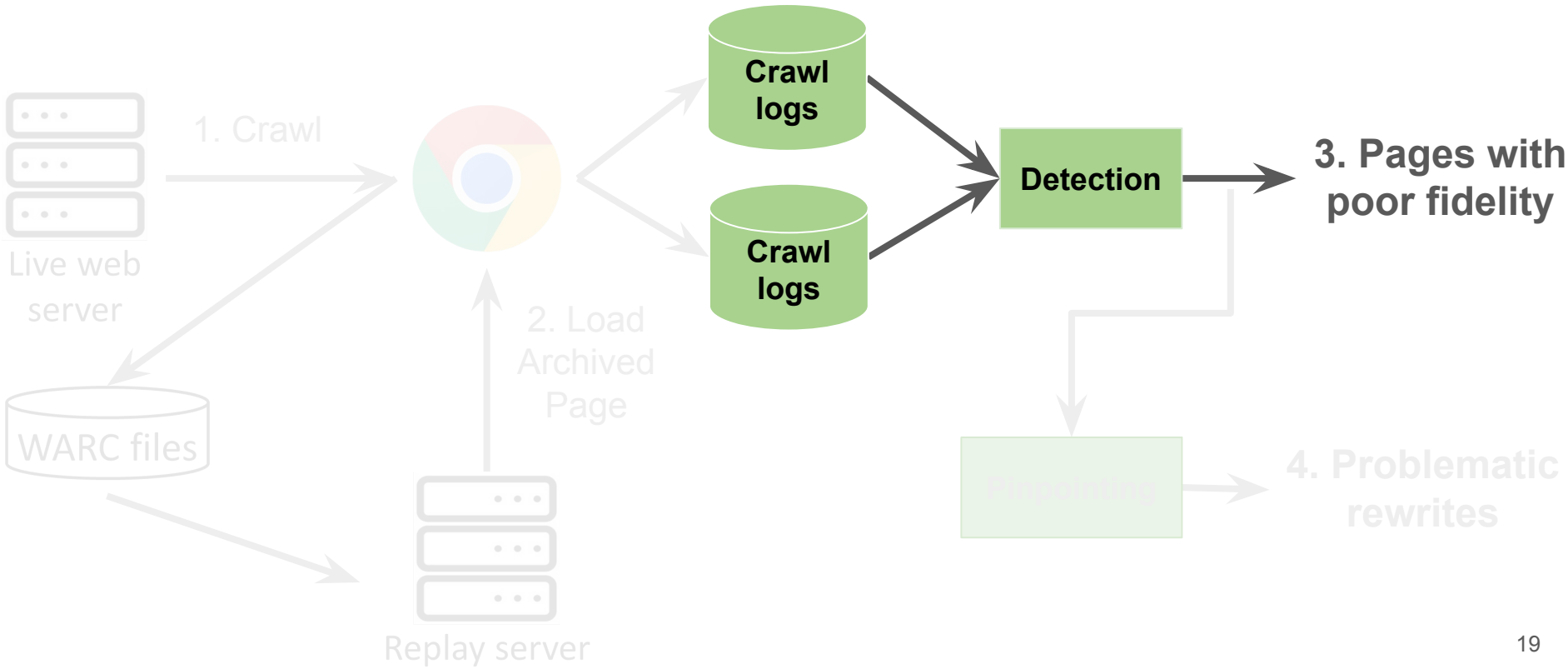
Fidelity violations ⇏ Runtime errors

Semantic errors, no error messages

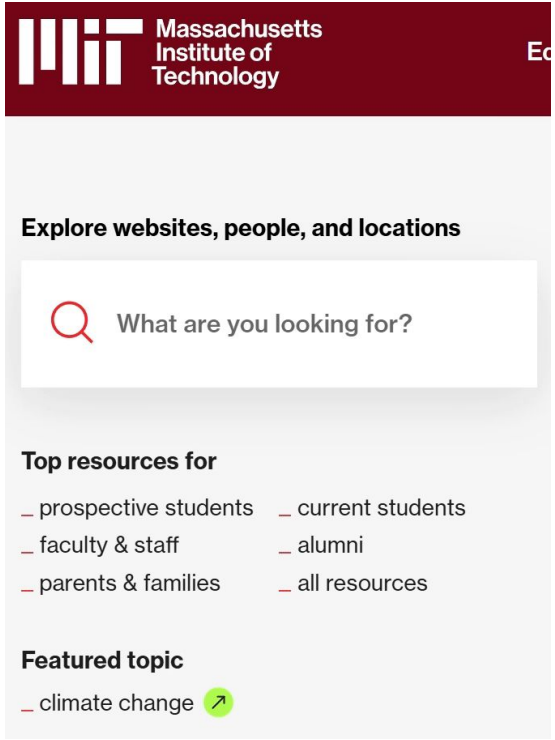




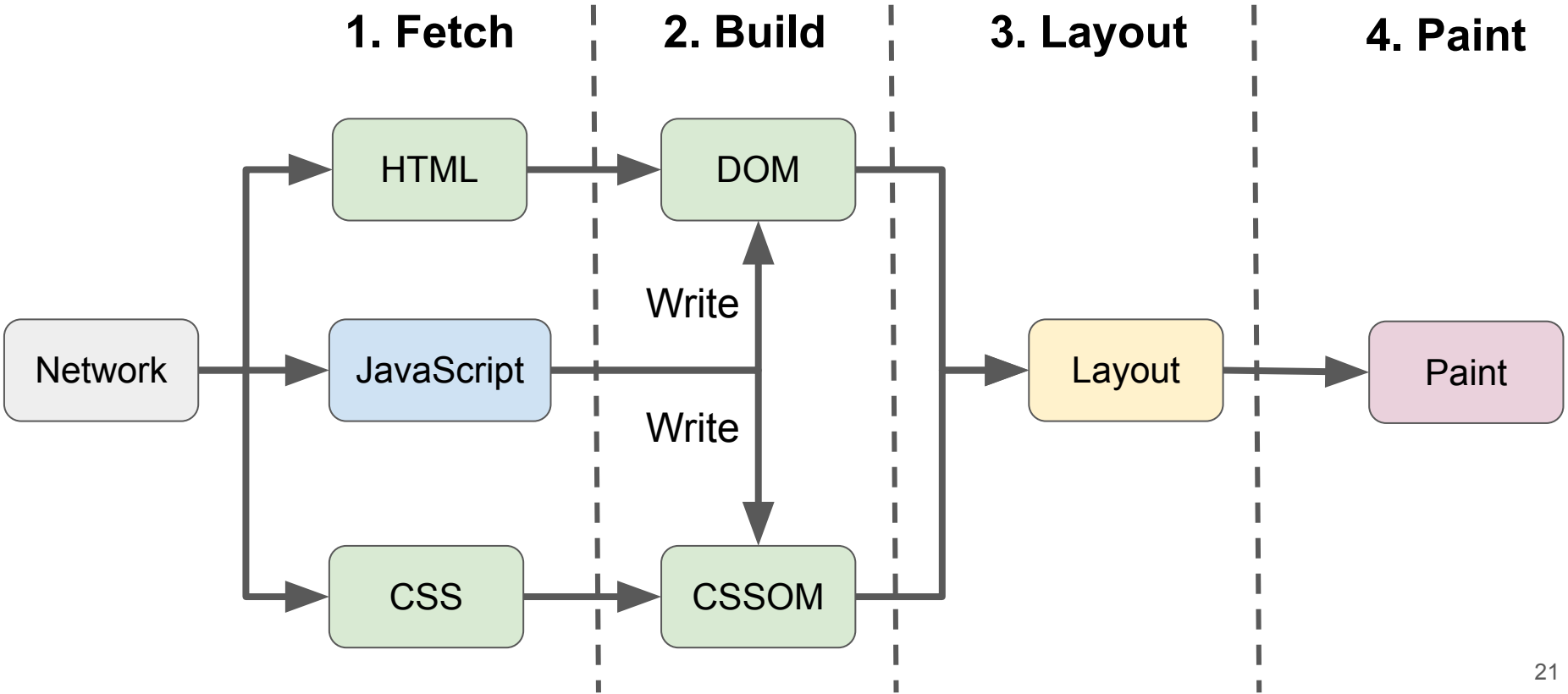
FidEx (**Fid**elity violation **Ex**poser)



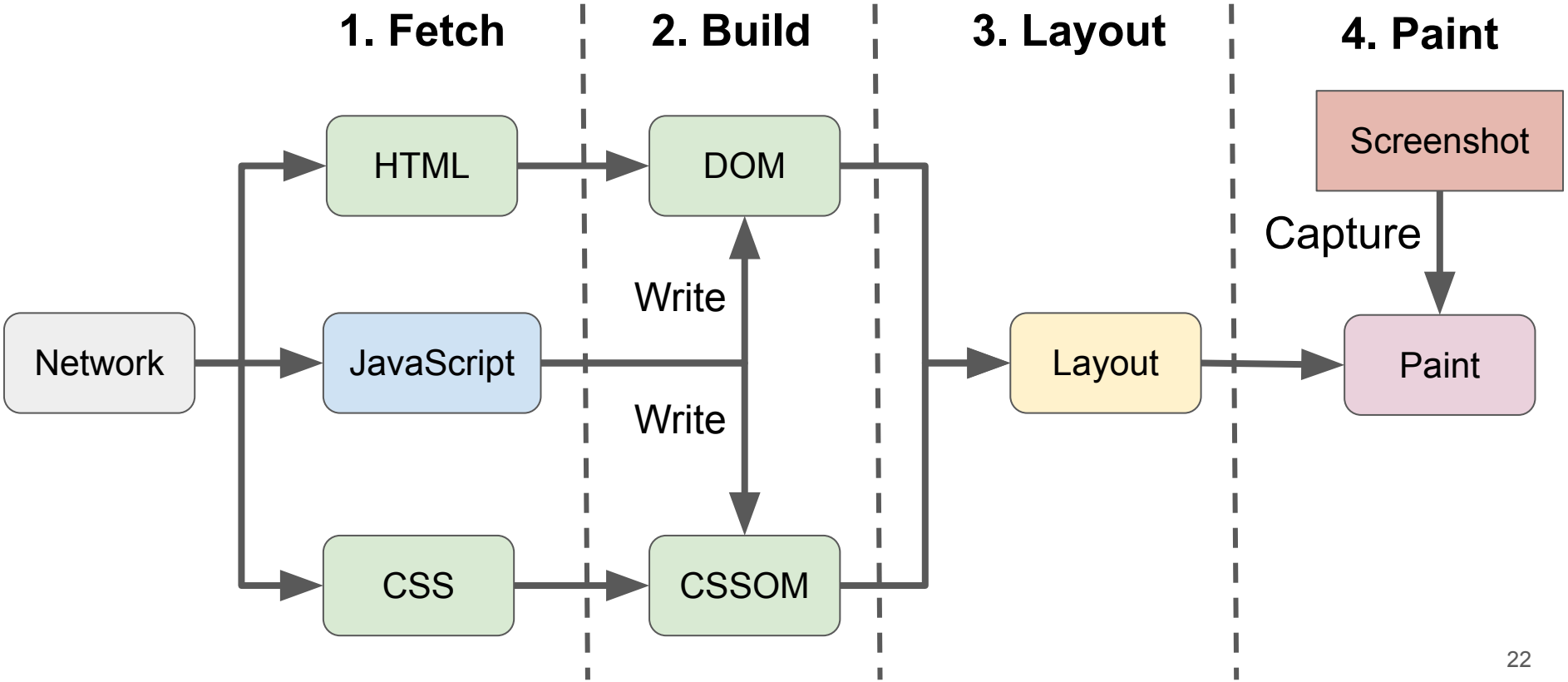
Screenshots can differ across multiple loads of a page



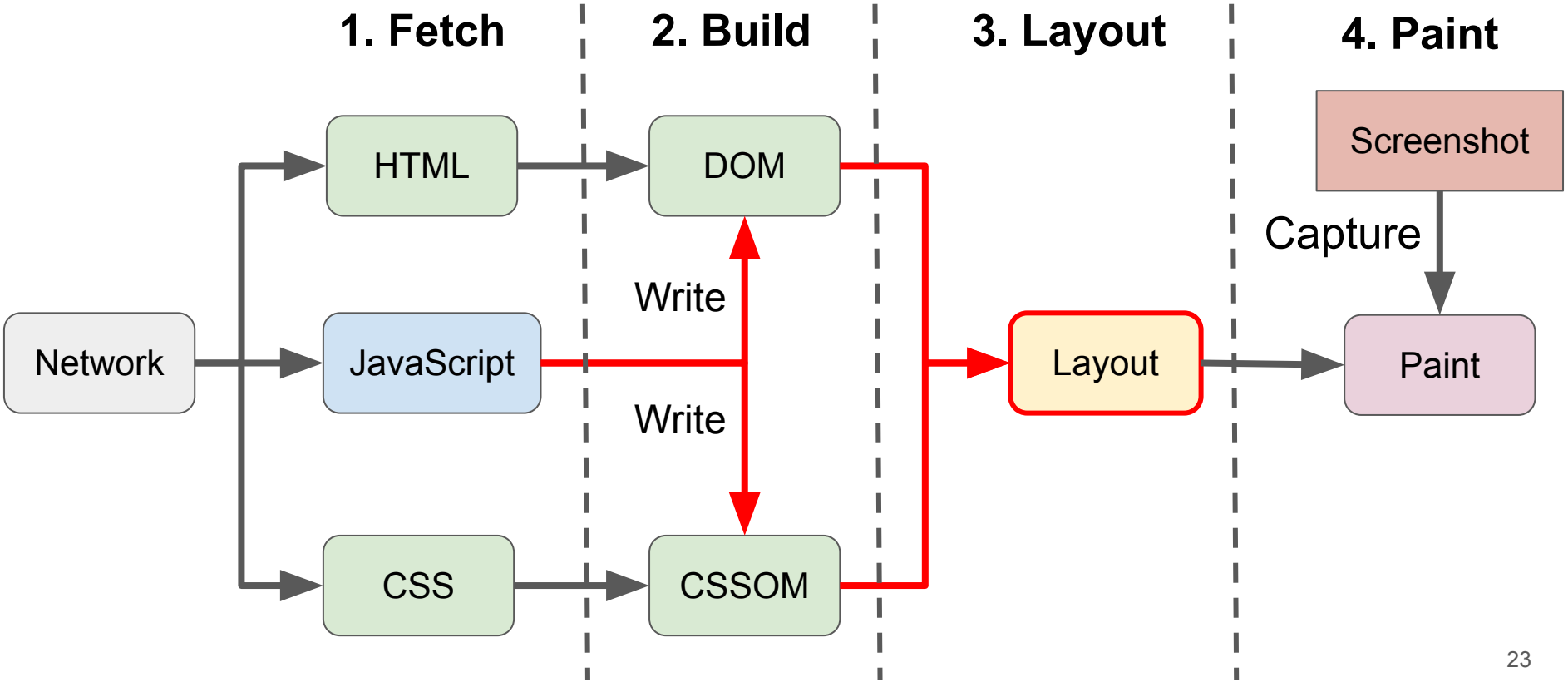
Critical rendering path in web browsers



Screenshot captures (constantly changing) visual output

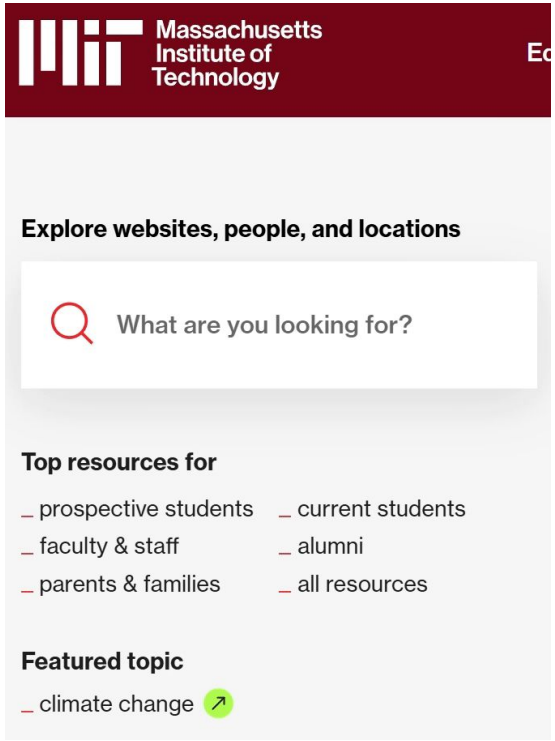


Layout and JS writes better represent a page

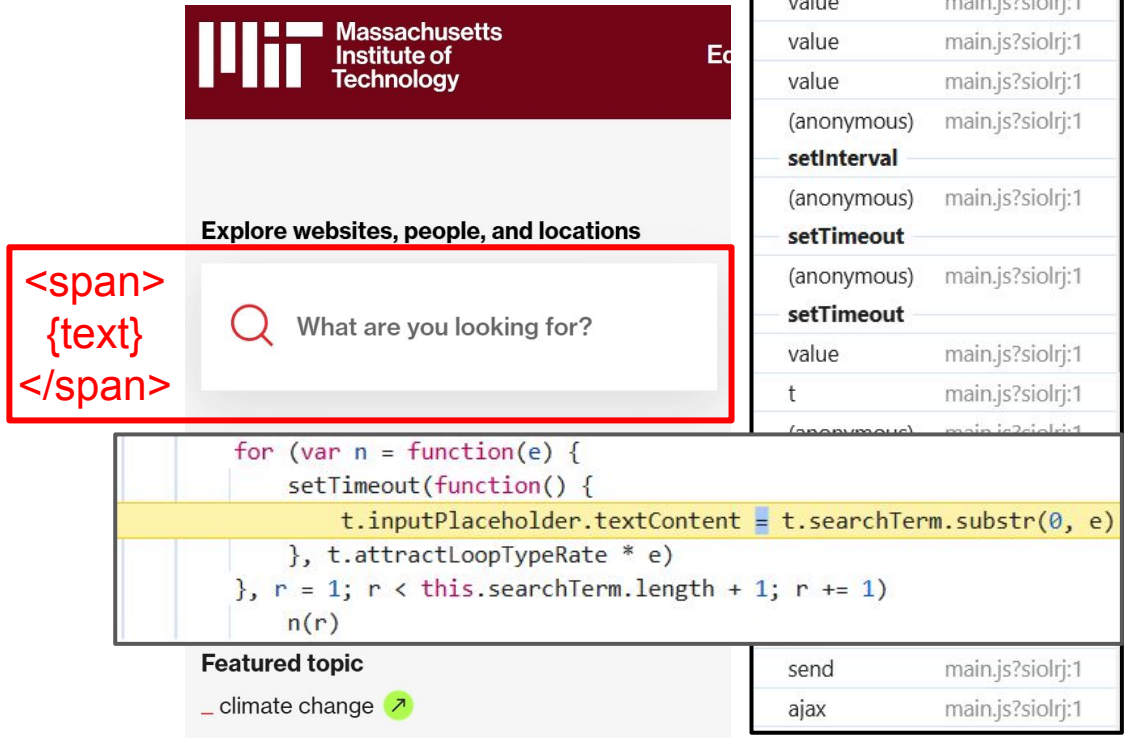


Fidelity violation detection - match layout tree

Layout tree unchanged



Fidelity violation detection - match JS writes

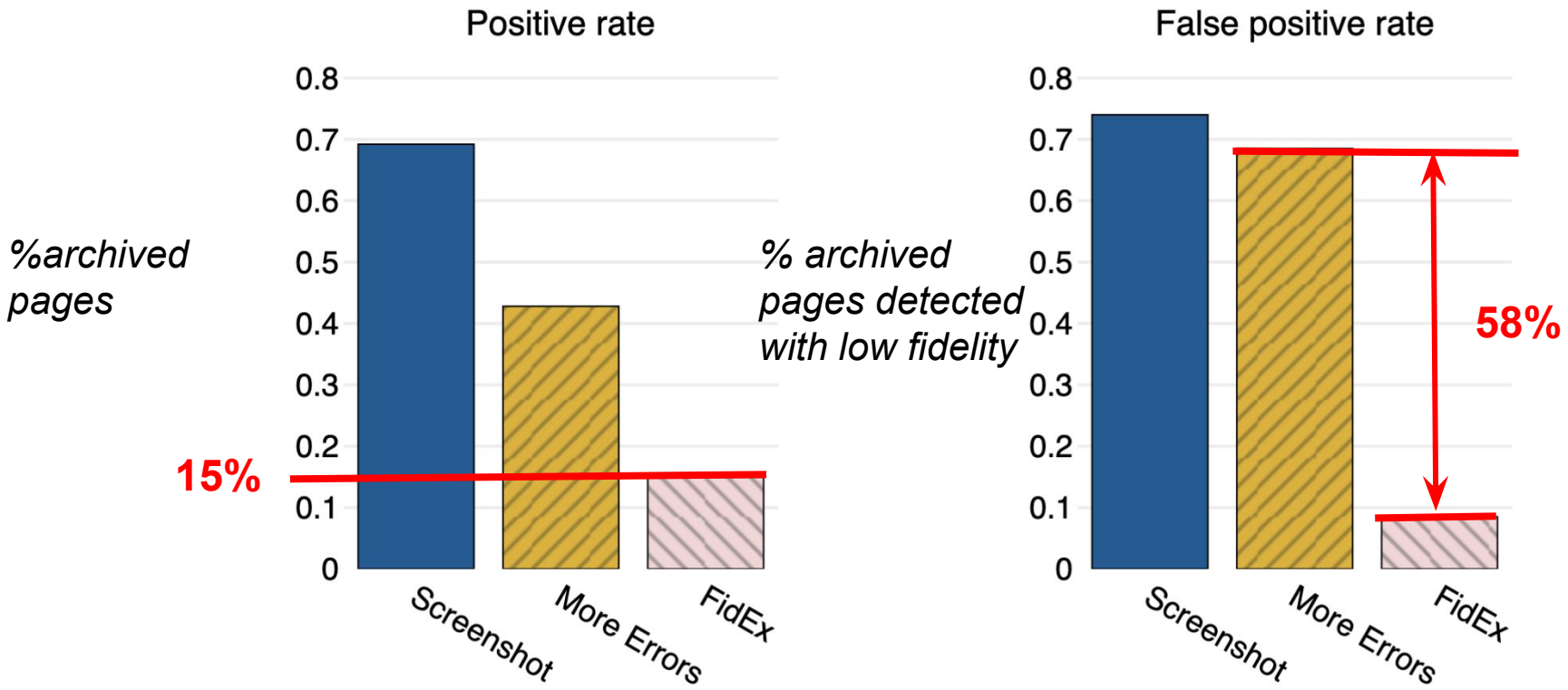


Associated {JS writes} unchanged

Evaluation

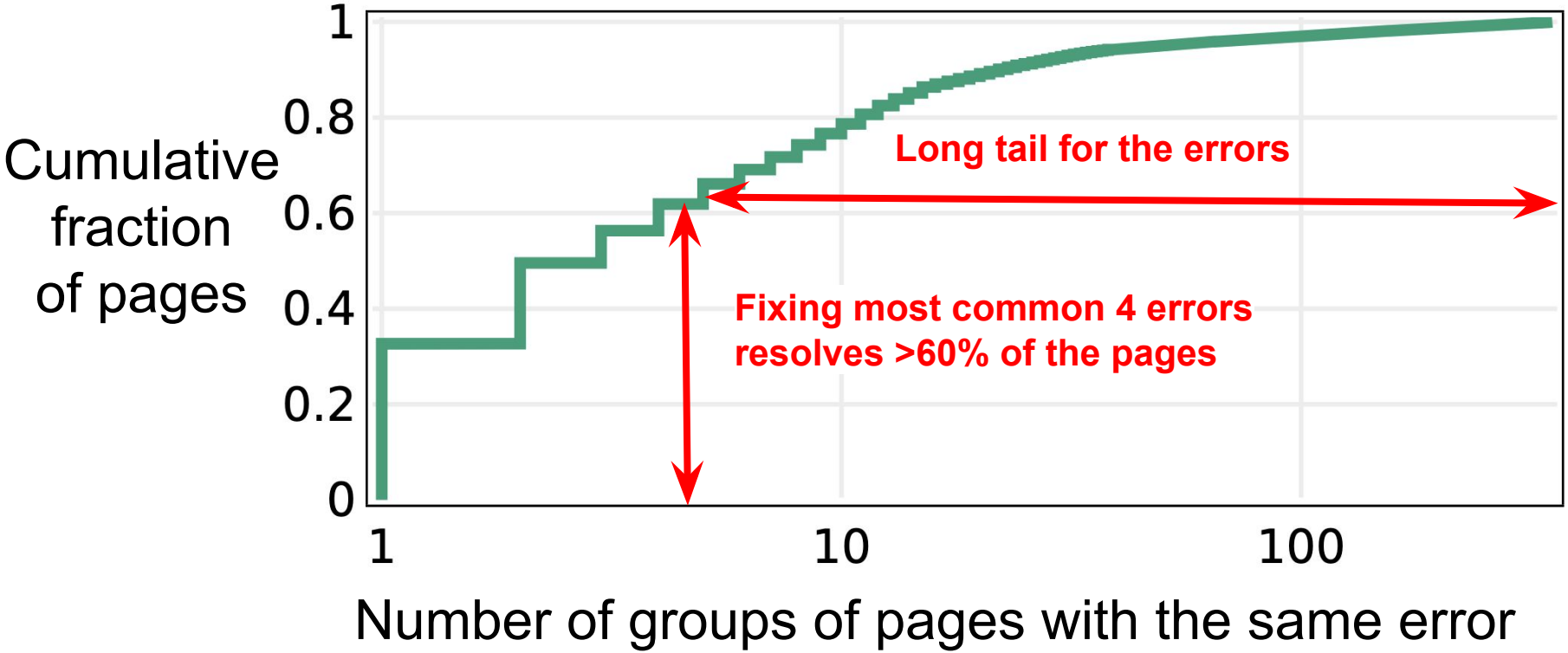
- **80K pages**
  - Sample from top 1 million sites, 5 pages per site
- **Detection baselines**
  - Match screenshot (Browsertrix's QA)
  - If archive has extra errors (More Errors)
- **Diagnosis baseline**
  - Pinpoint extra errors in archived copy

FidEx improves accuracy of detecting fidelity violations



Long-tail exists, but common errors dominate

Long-tail exists, but common errors dominate



Research → Impact

- Developed fixes in pywb for 4 most common errors
  - **~64%** of target pages fully fixed
  - **No new page** is broken

- Reduced %fidelity violations from **15%** to **9%**

Summary

- Replay on archived copies needs to rewrite JS
- Erroneous rewrites lead to fidelity violations
- We design and implement FidEx
  - Automatically detects fidelity violations accurately
  - Pinpoints erroneous rewrites for developers to fix

Thank you!

https://github.com/USC-NSL/FidEx