

Detecting and Diagnosing Errors in Replaying Archived Web Pages

Jingyuan Zhu

Huanchen Sun

Harsha V. Madhyastha



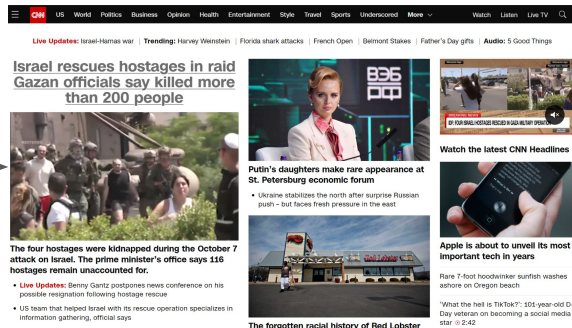
USC University of
Southern California

Web archives capture and serve snapshots of webpages

Record:

Crawler

Crawl



Generate

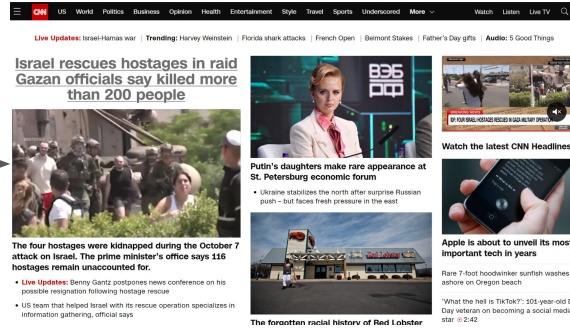


Web archives capture and serve snapshots of webpages

Record:

Crawler

Crawl



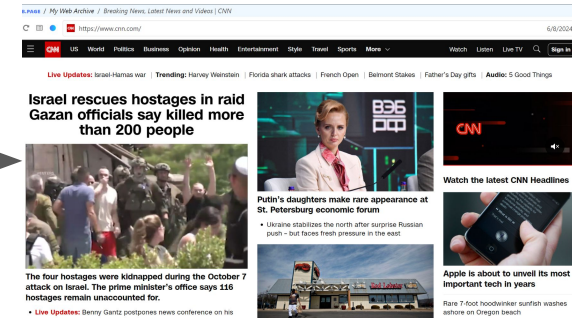
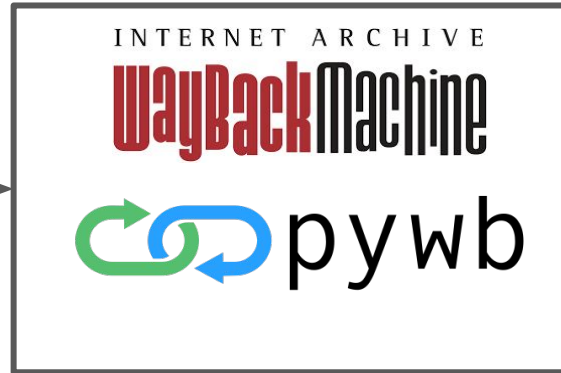
Generate



Replay:

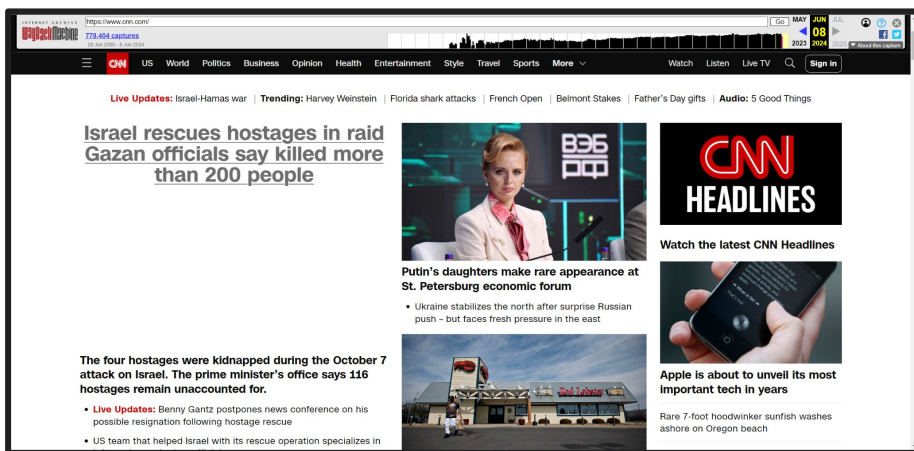
.warc

Input

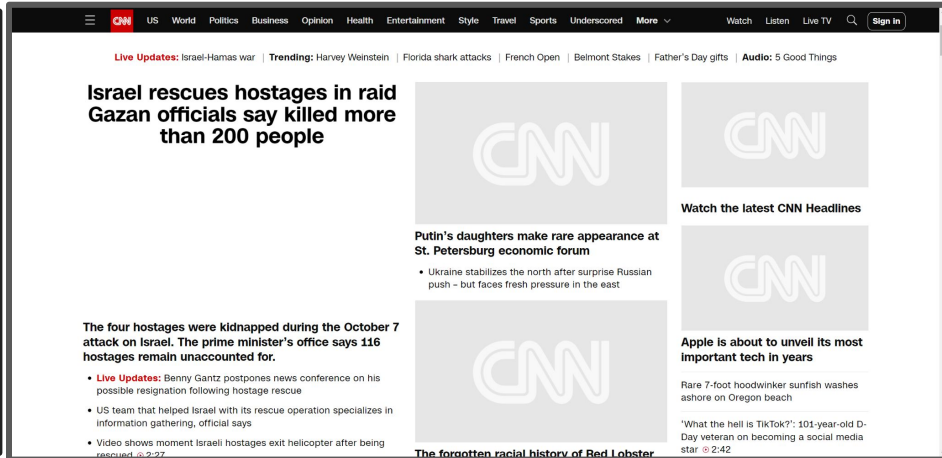


Need to rewrite URLs for crawled resources

With rewriting

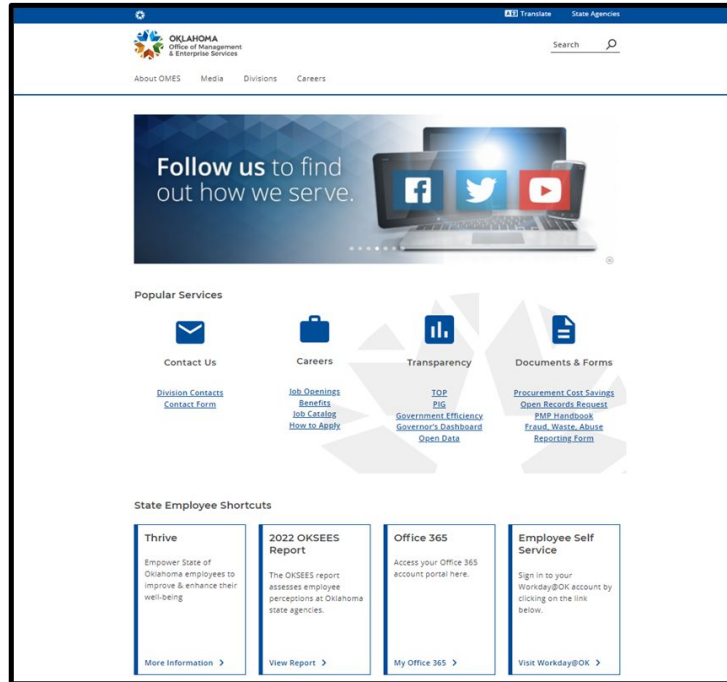


Without rewriting

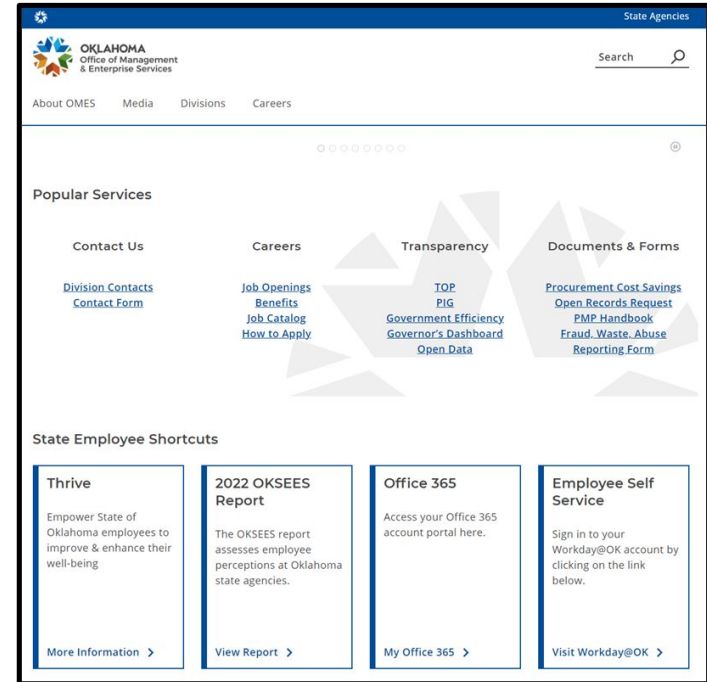


However, JS rewrites can lead to fidelity issues

Live web page

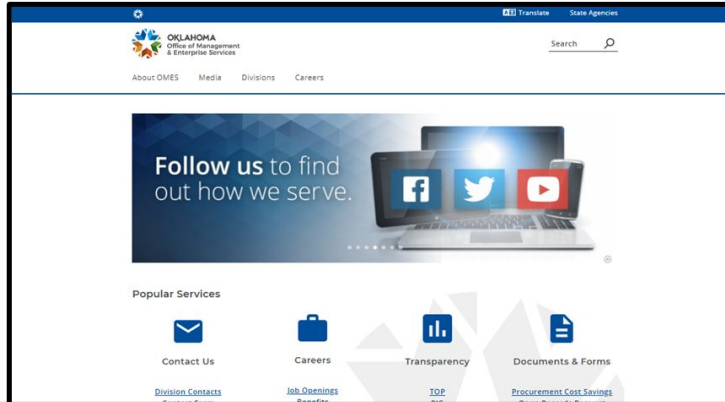


Archived copy

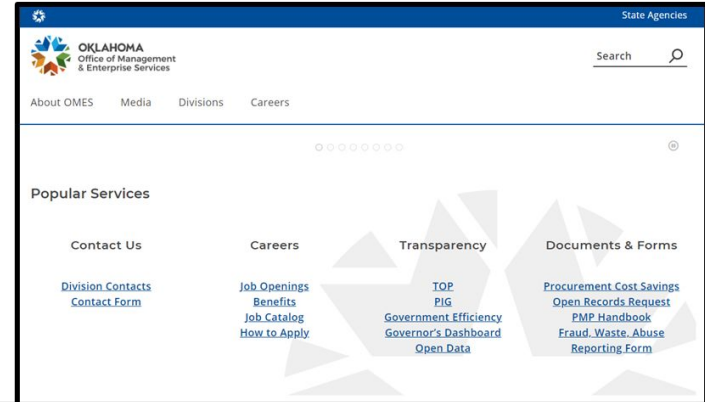


However, JS rewrites can lead to fidelity issues

Live web page



Archived copy



```
✖ Uncaught TypeError: string.indexOf is not a function
    at Wombat.startsWithOneOf (wombat.js:21:21211)
    at Object.get (wombat.js:21:61588)
    at readData (clientlib-base.min.js:2450:22)
    at HTMLDocument.onDocumentReady (clientlib-base.min.js:2695:56)
```

wombat.js:21

[More Information >](#)

[View Report >](#)

[My Office 365 >](#)

[Visit Workday@OK >](#)

[More Information >](#)

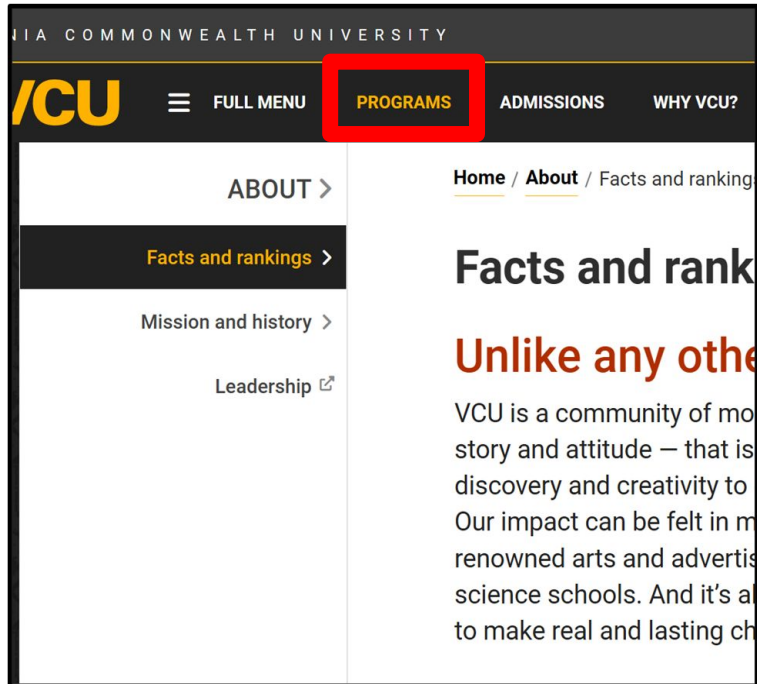
[View Report >](#)

[My Office 365 >](#)

[Visit Workday@OK >](#)

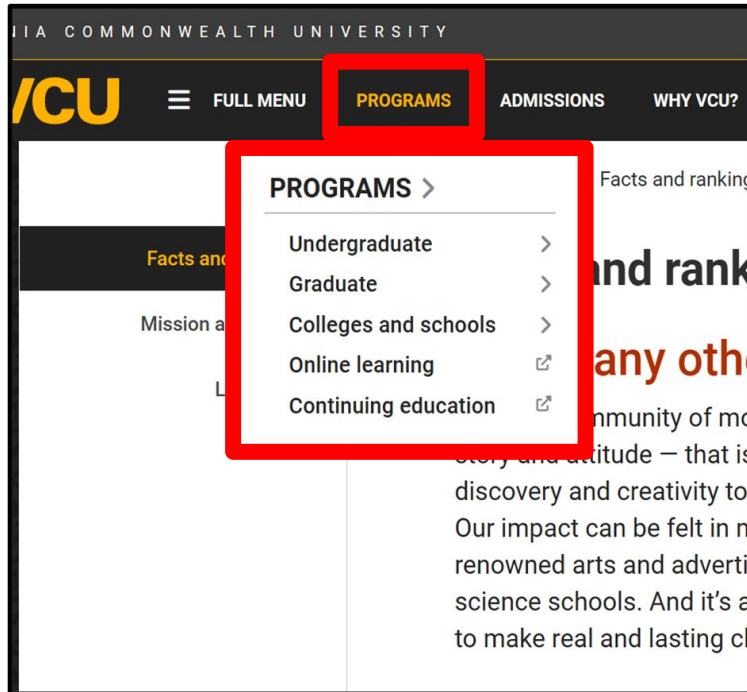
However, JS rewrites can lead to fidelity issues

Live web page



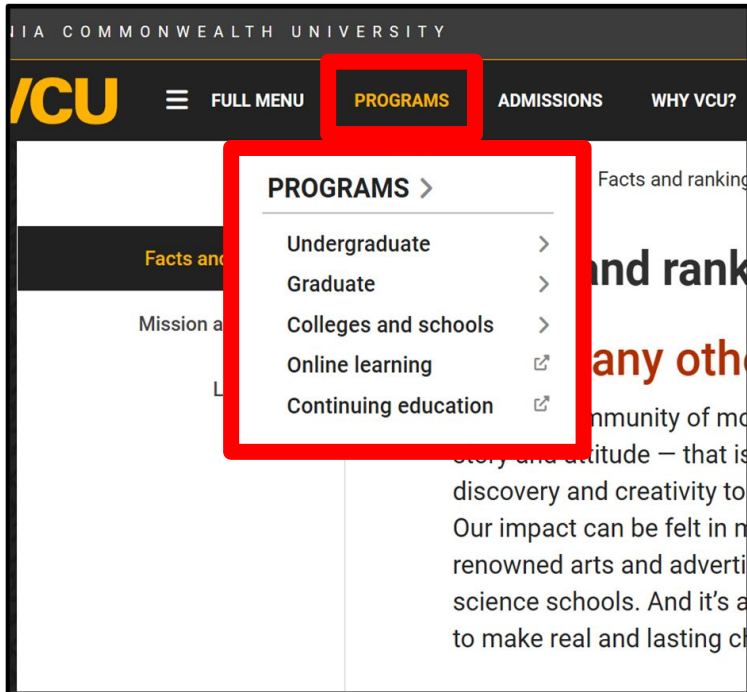
However, JS rewrites can lead to fidelity issues

Live web page

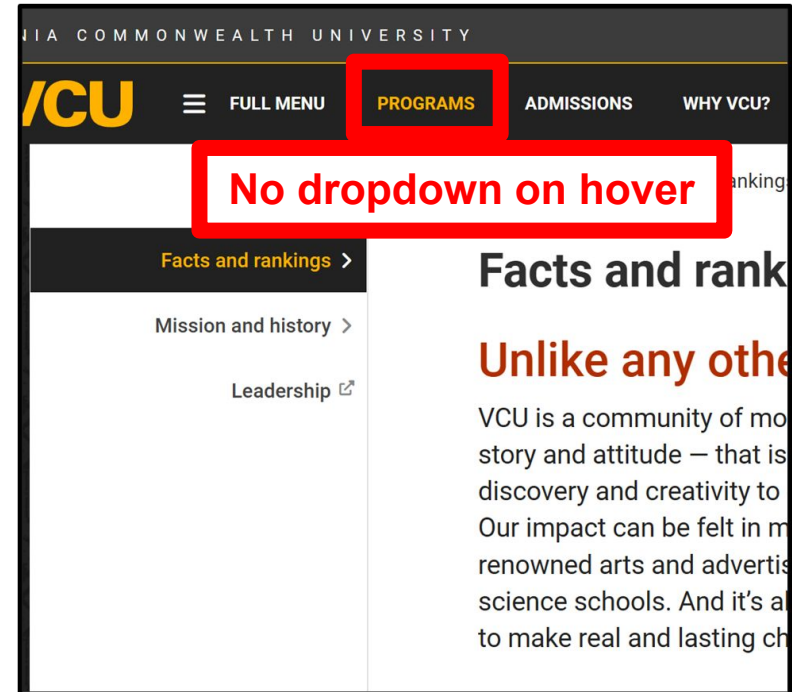


However, JS rewrites can lead to fidelity issues

Live web page

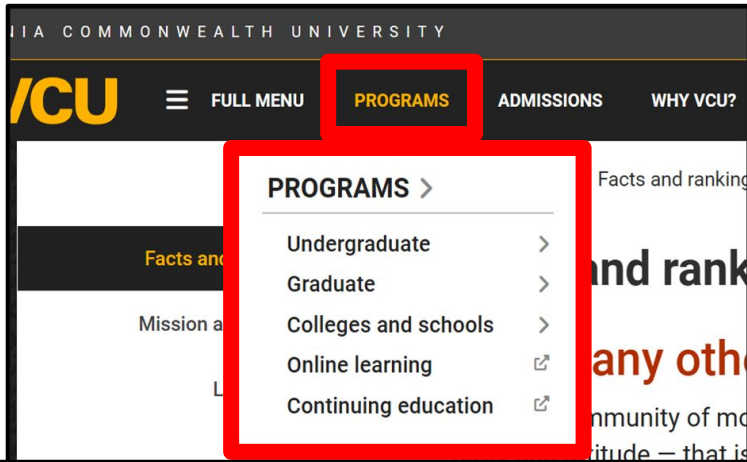


Archived copy

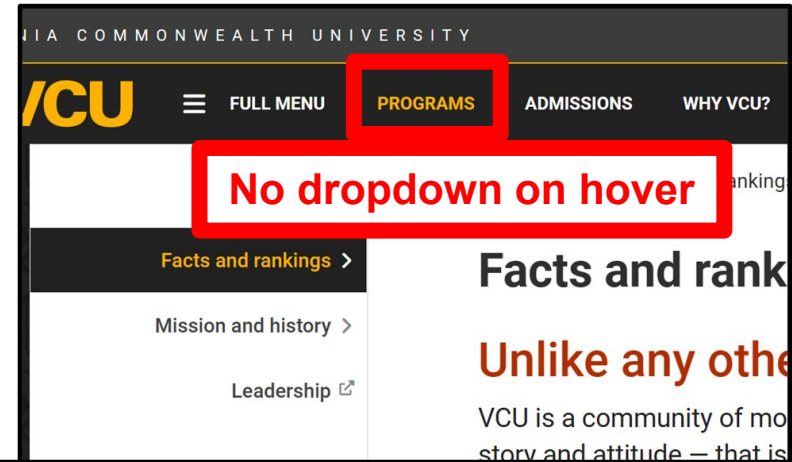


However, JS rewrites can lead to fidelity issues

Live web page



Archived copy



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


Goals of our project


- Detect fidelity violations when replaying page snapshots
 - Focus only on fidelity violations caused due to rewriting
 - We do not address incomplete/inaccurate recording


Goals of our project


- Detect fidelity violations when replaying page snapshots
 - Focus only on fidelity violations caused due to rewriting
 - We do not address incomplete/inaccurate recording
- Diagnose erroneous rewrites which lead to fidelity issues
 - Not trying to *fix* fidelity issues
 - Once diagnosed, developers will need to identify necessary fixes


Existing Quality Assurance in Web Archiving

 Screenshots **45%**

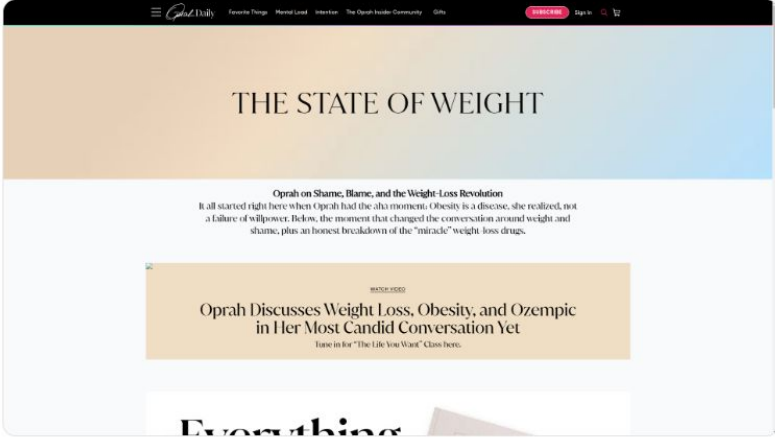
 Text **100%**

 Resources

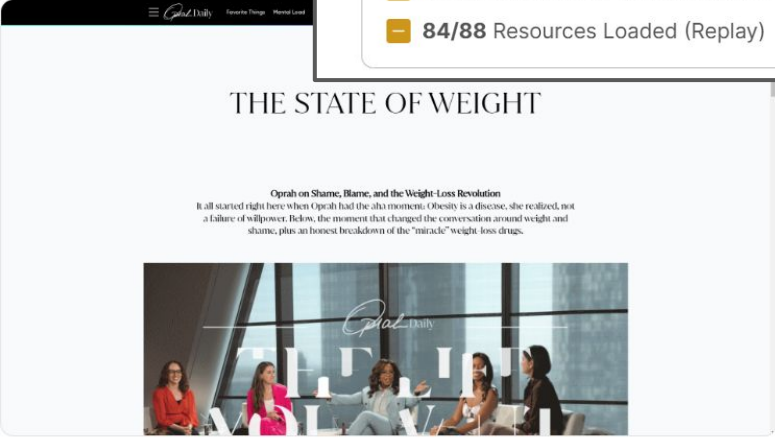
 Replay


 <https://www.oprahdaily.com/stateofweight/>


Screenshot during crawl





Screenshot during analysis




 **The State of Weight**
<https://www.oprahdaily.com/stateofweight/>

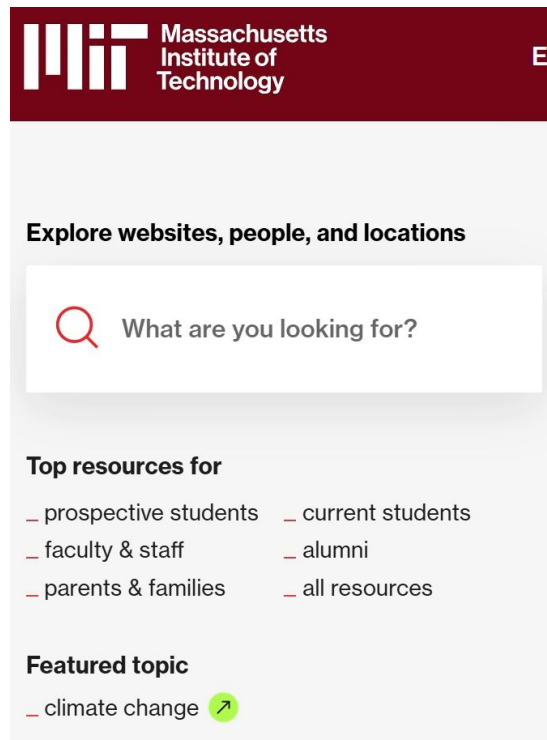
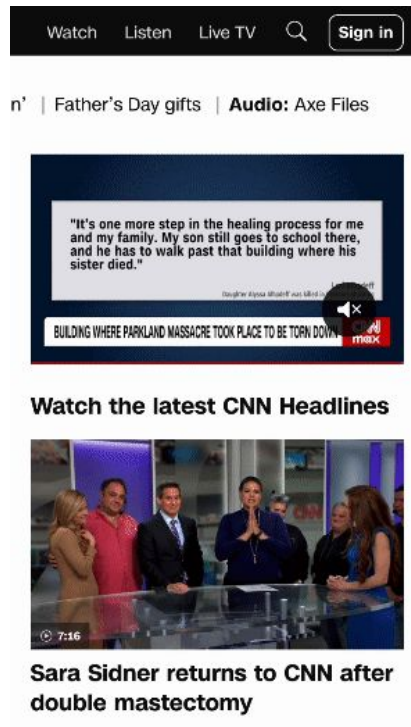
 **45.33%** Screenshot Match

 **100%** Extracted Text Match

 **113/114** Resources Loaded (Crawl)

 **84/88** Resources Loaded (Replay)

Screenshots Differ across Multiple Loads of Same Page



Runtime Error \nRightarrow Fidelity Violation

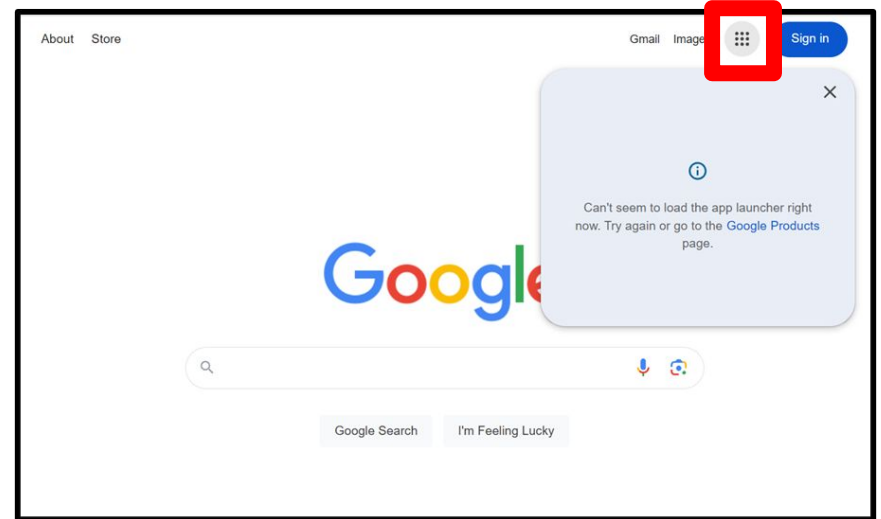
Hard to analyze if an error will cause violations
by manually inspecting its stack trace

▼ Call Stack	
➤ m	metroRequire.f0...s:sourcemap:14
d	metroRequire.f0...s:sourcemap:14
i	metroRequire.f0...s:sourcemap:14
m.exports	
	asyncRequire.b6...s:sourcemap:13
c.scheduler.postTask.event	
	initializer.3de...s:sourcemap:17
postTask	
s	5807.0bd3811f6a...:sourcemap:204
await in s	
e.default	
	initializer.3de...s:sourcemap:17
n.consumerId	
	initializer.3de...s:sourcemap:28
(anonymous)	
	f29e.9df4fc1466...s:sourcemap:55
Promise.then	
instanceCreator	
	f29e.9df4fc1466...s:sourcemap:55
(anonymous)	
	f29e.9df4fc1466...s:sourcemap:55
createInstances	
	f29e.9df4fc1466...s:sourcemap:55
e.createInitializerInjector	
	5807.0bd3811f6a...:sourcemap:193

Fidelity Violation \nRightarrow Runtime Error

Errors can be **hard to detect**

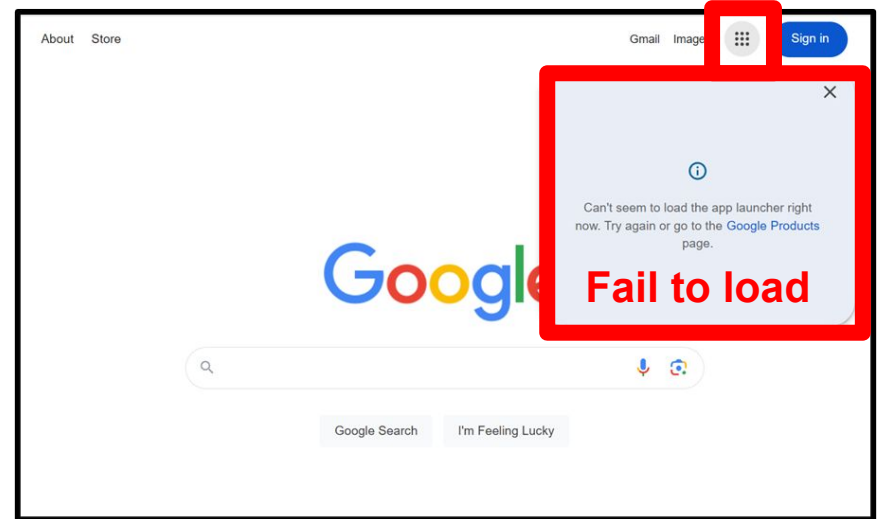
- Semantic errors, no error messages



Fidelity Violation \nRightarrow Runtime Error

Errors can be **hard to detect**

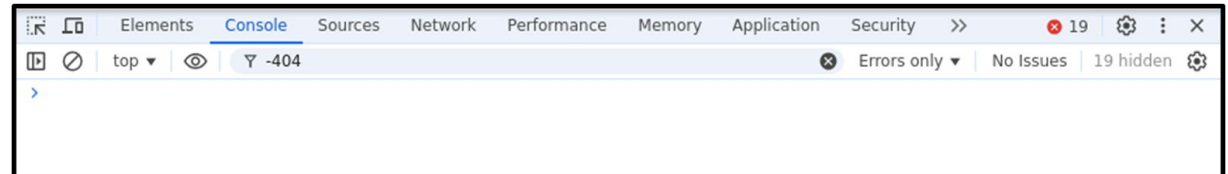
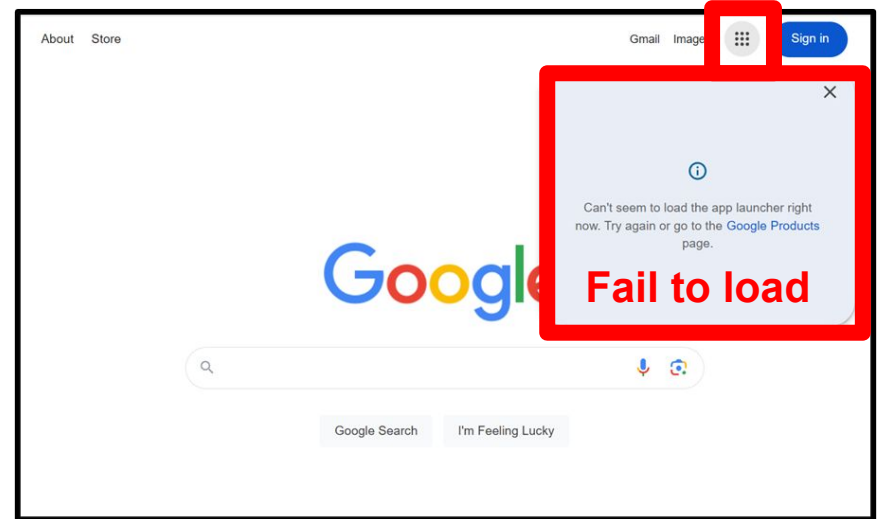
- Semantic errors, no error messages



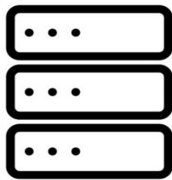
Fidelity Violation \nRightarrow Runtime Error

Errors can be **hard to detect**

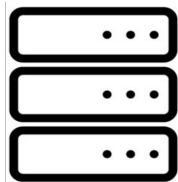
- Semantic errors, no error messages



FidEx (Fidelity violation **Ex**poser)

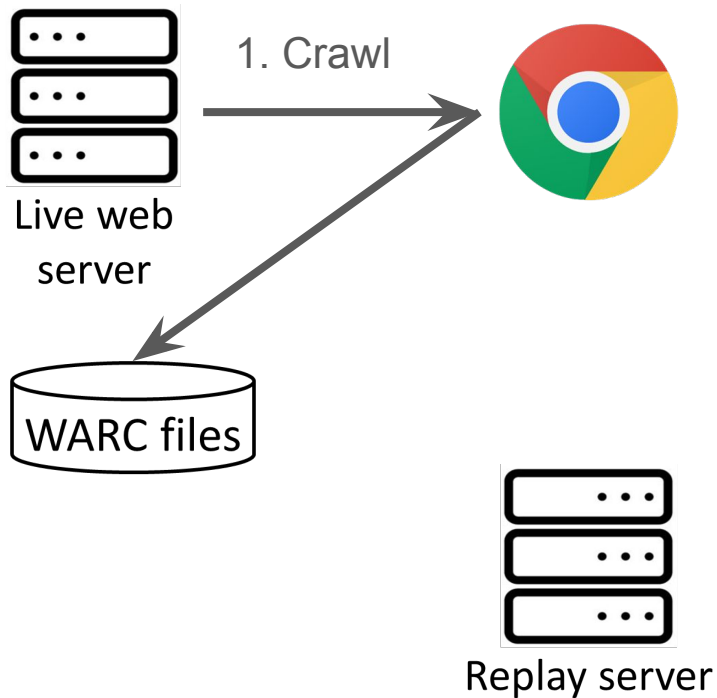


Live web
server

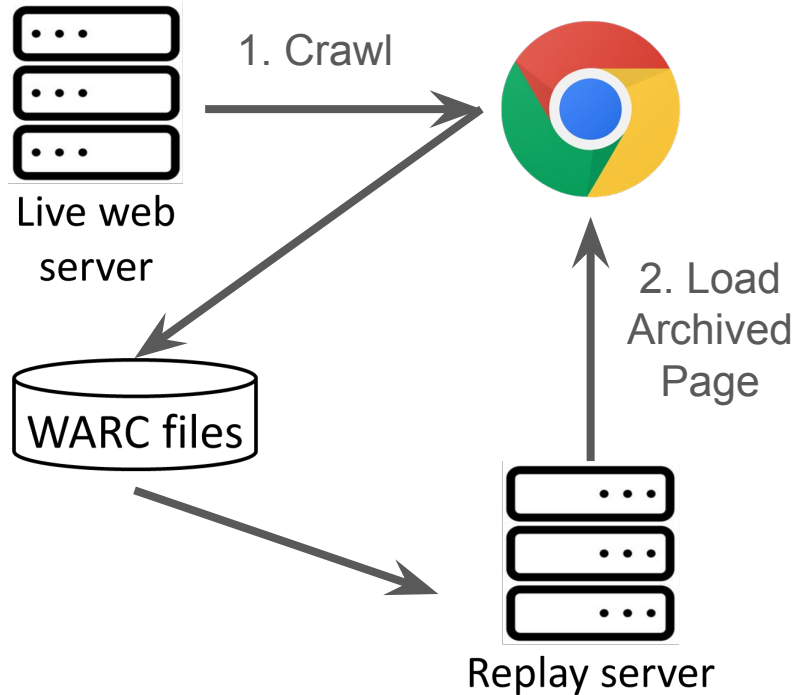


Replay server

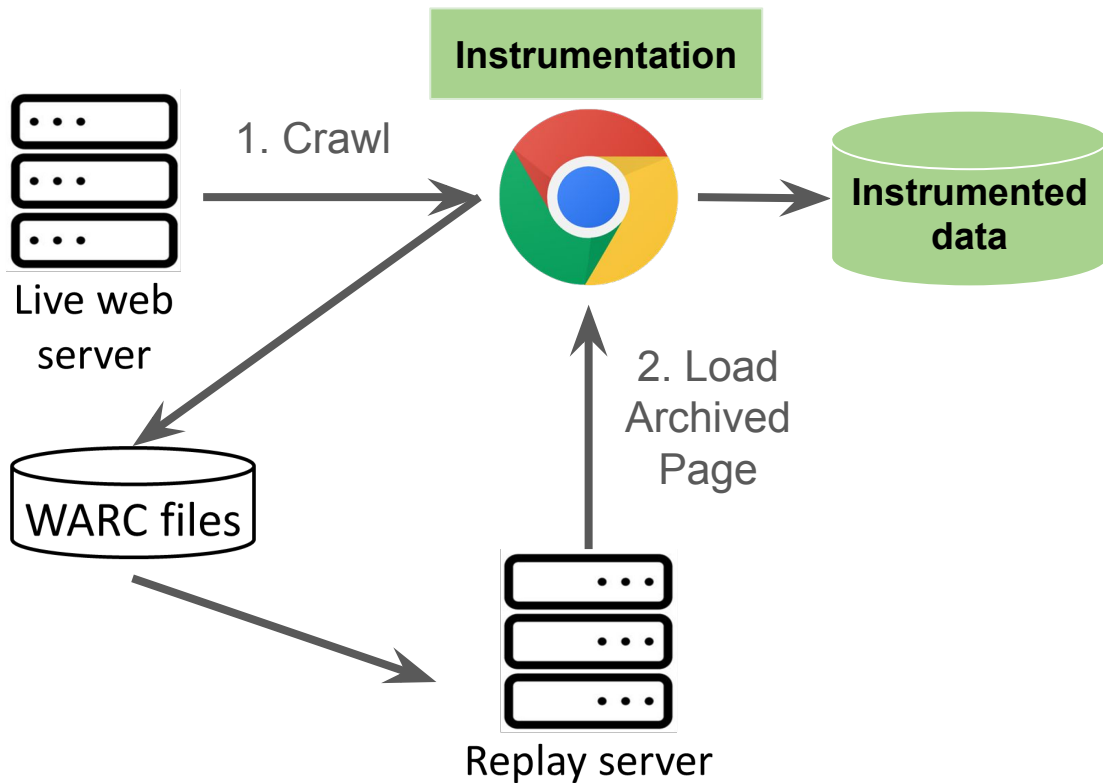
FidEx (Fidelity violation **Ex**poser)



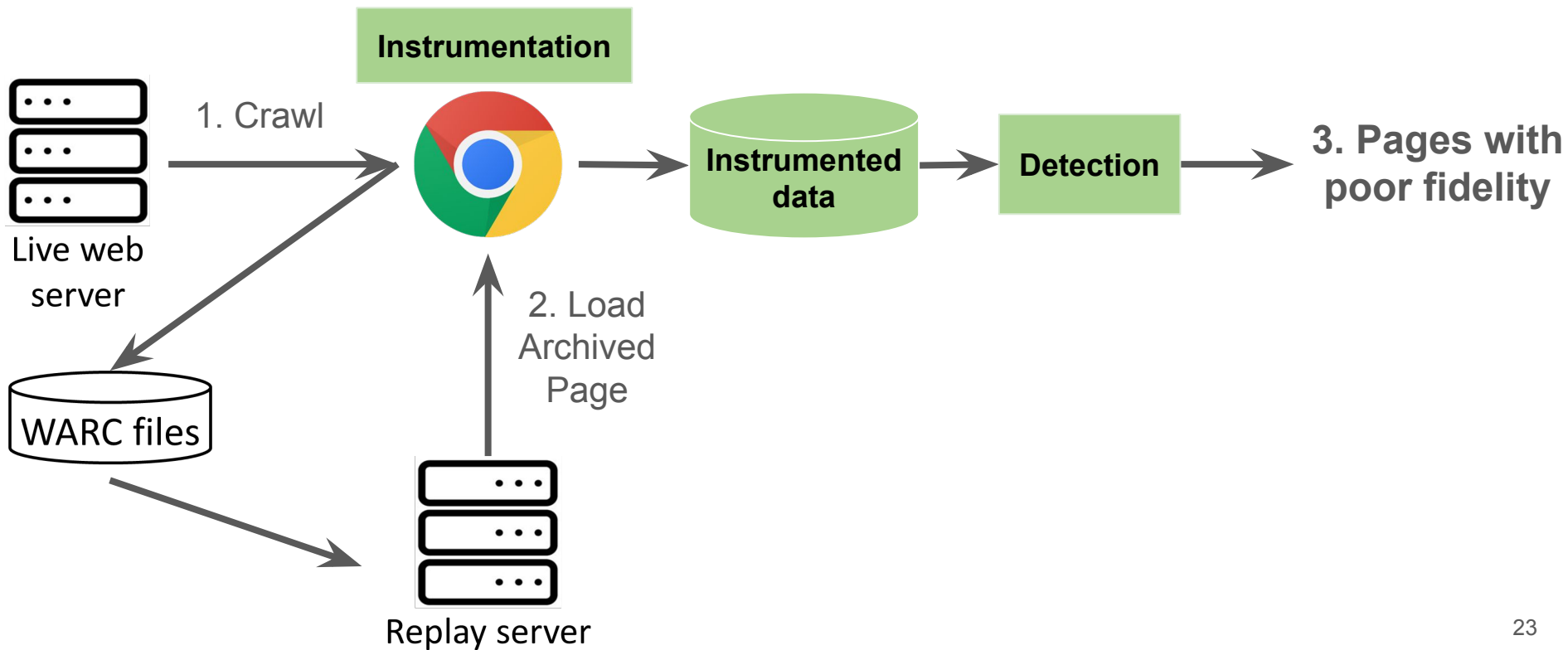
FidEx (Fidelity violation **Ex**poser)



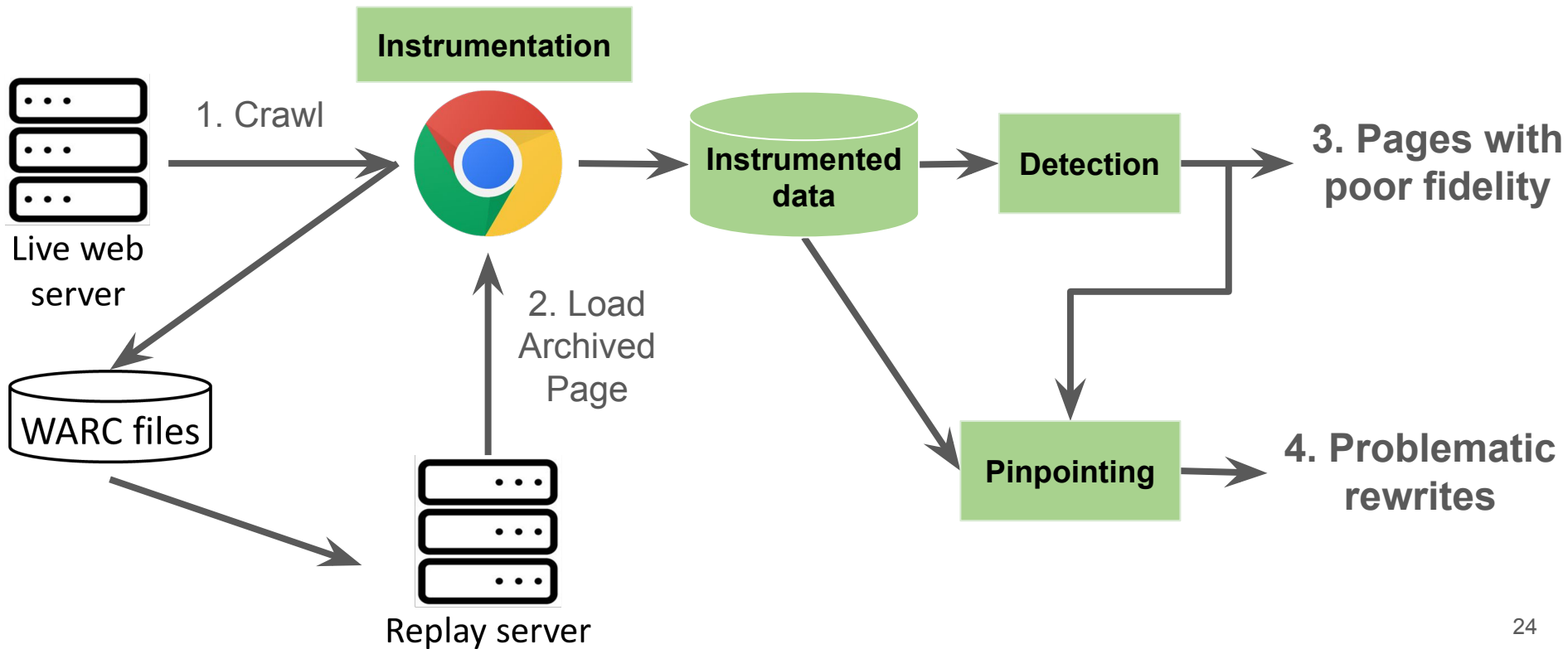
FidEx (Fidelity violation **Ex**poser)



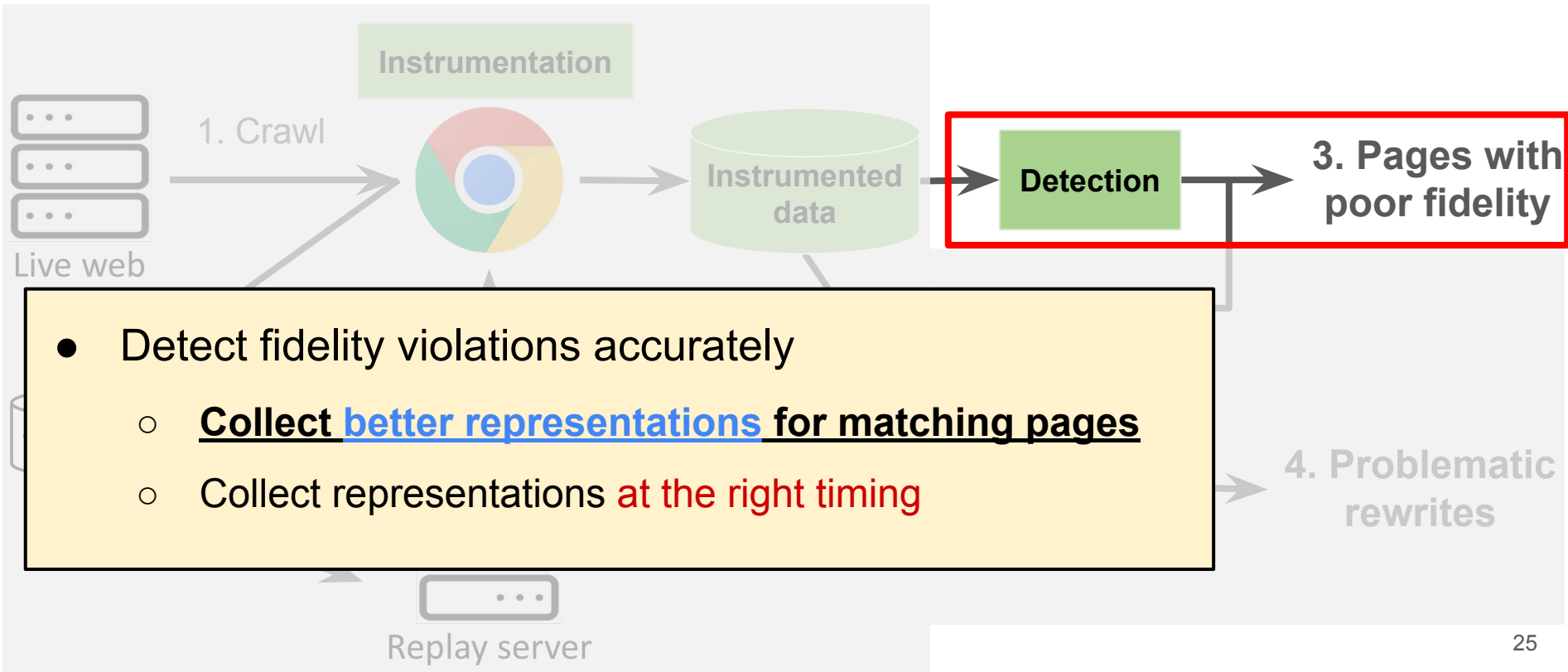
FidEx (Fidelity violation **Ex**poser)



FidEx (Fidelity violation **Ex**poser)



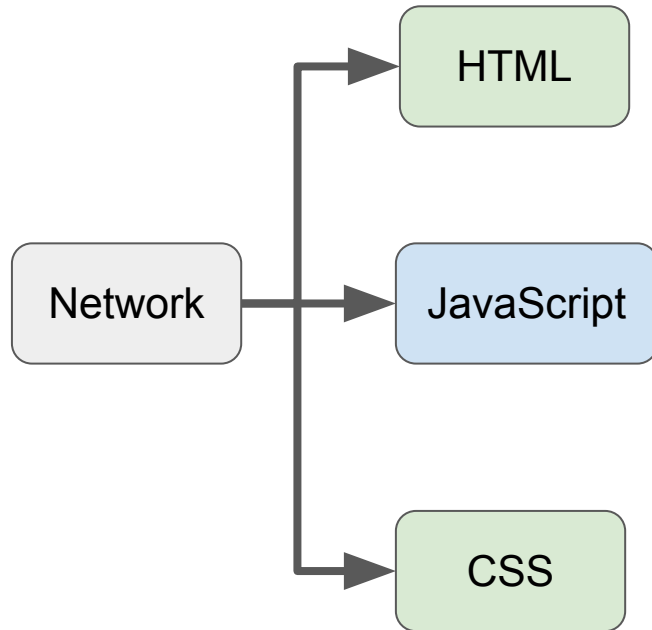
Fidelity violation detection



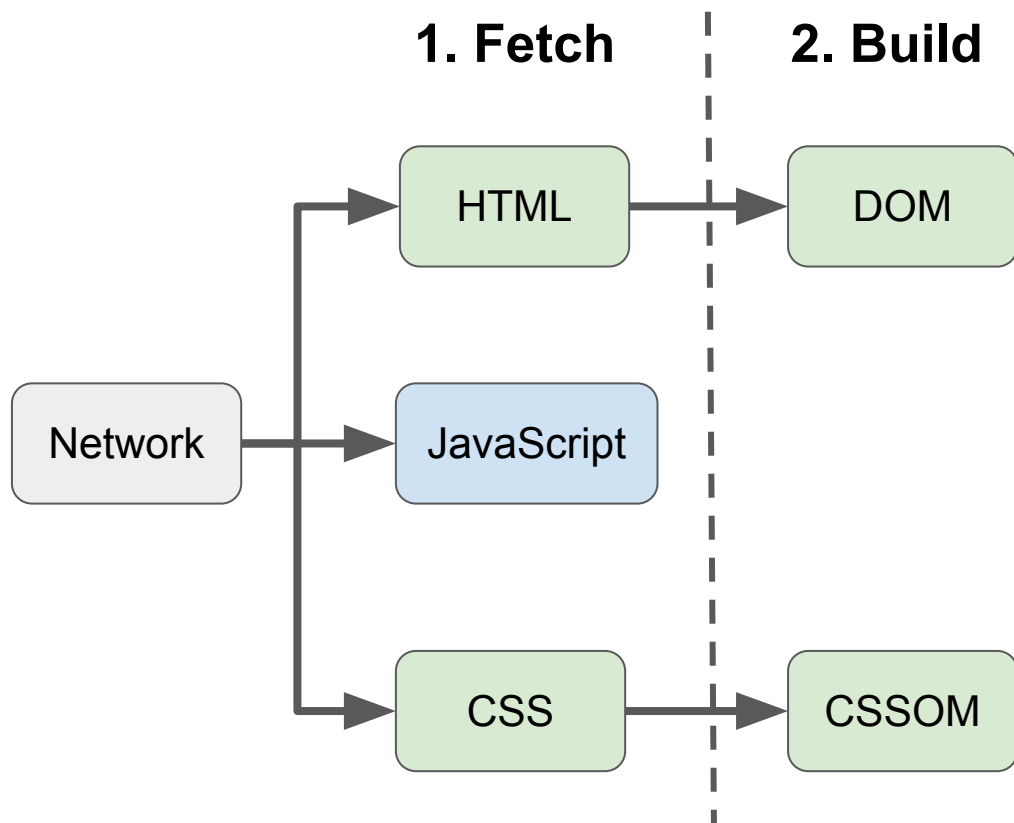
How a browser renders pages

How a browser renders pages

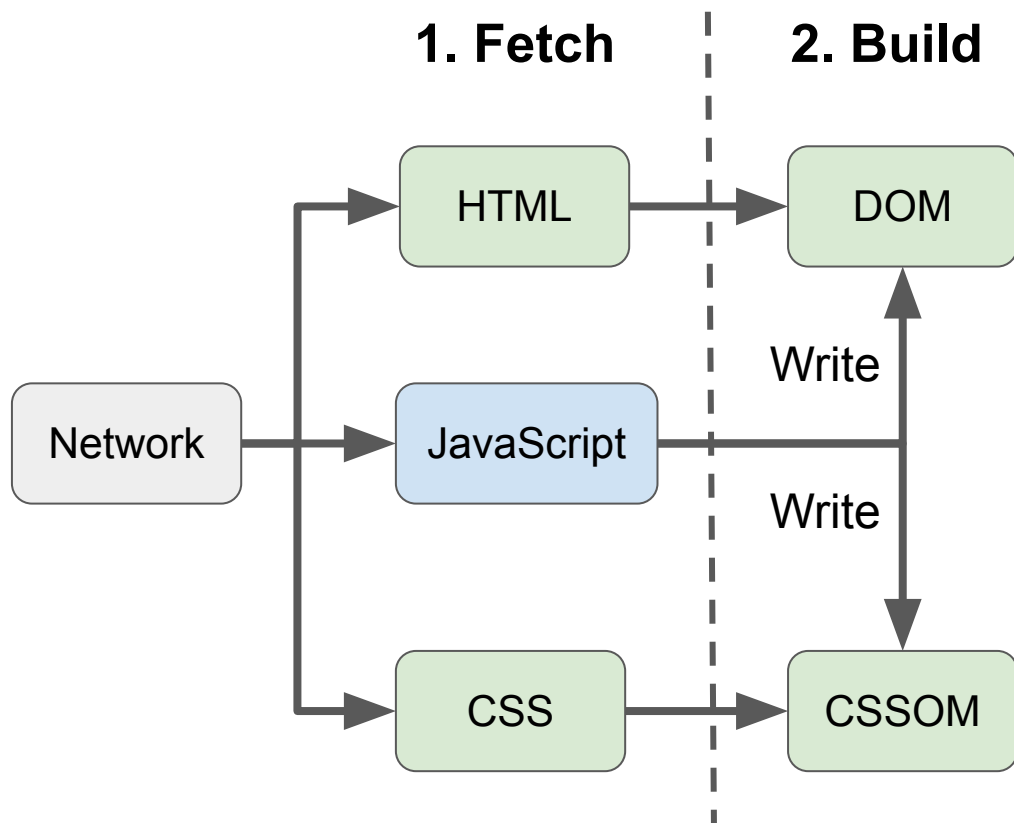
1. Fetch



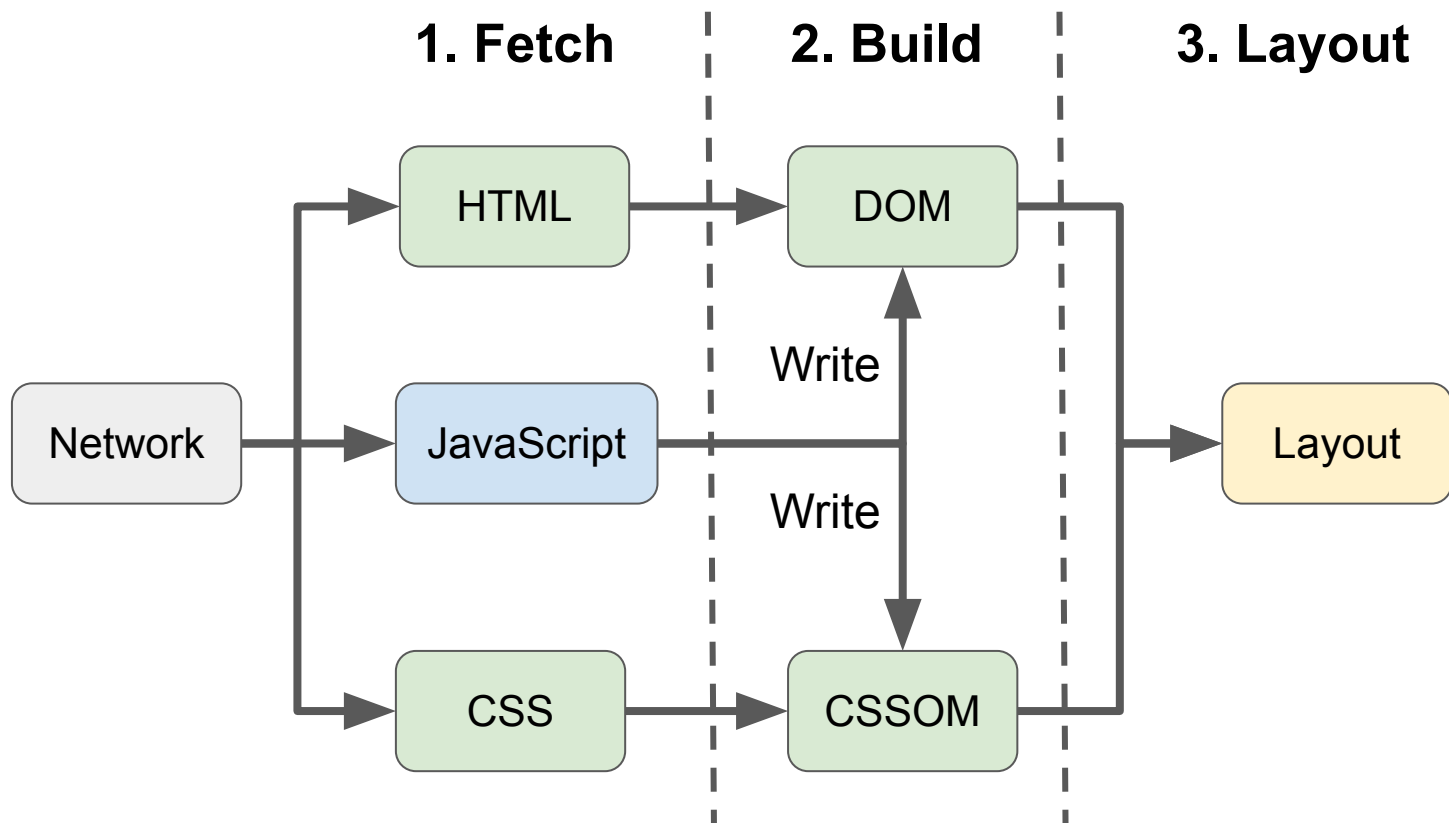
How a browser renders pages



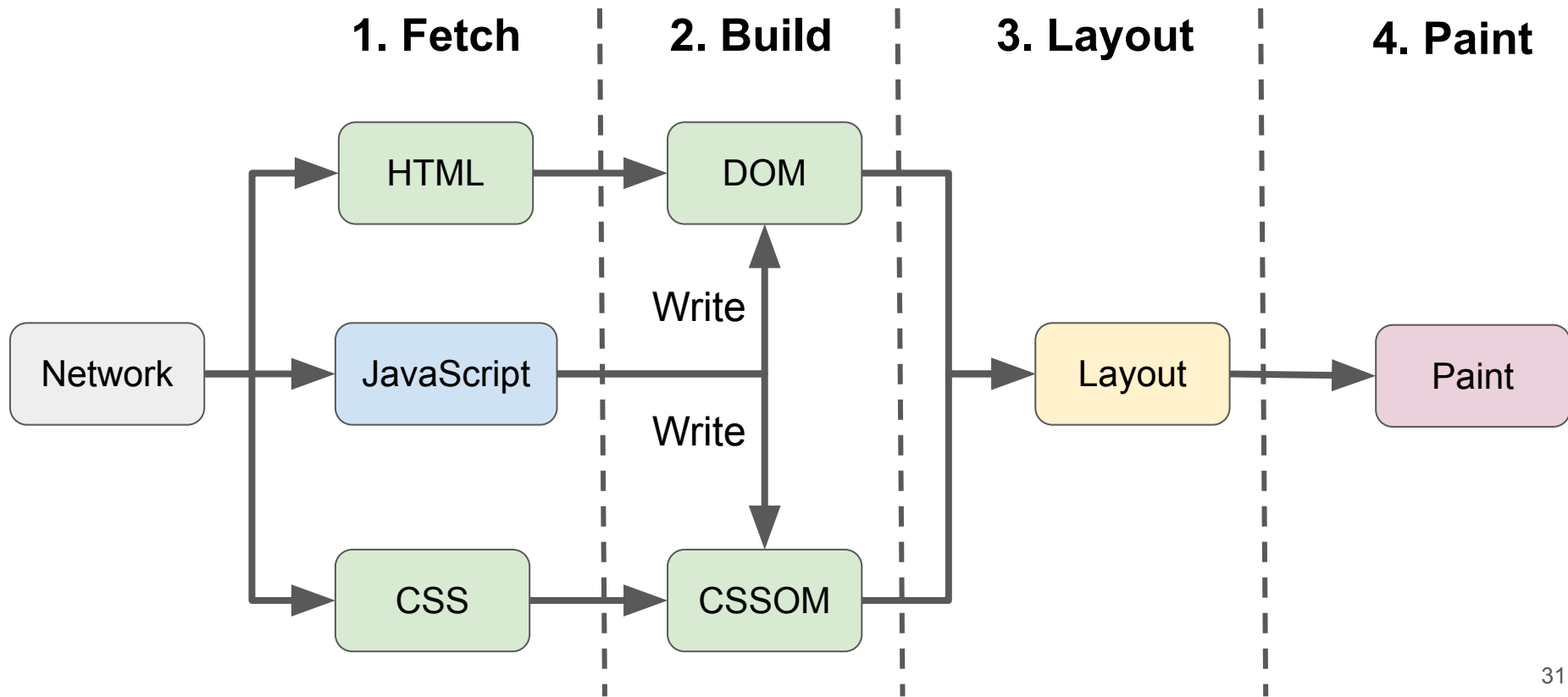
How a browser renders pages



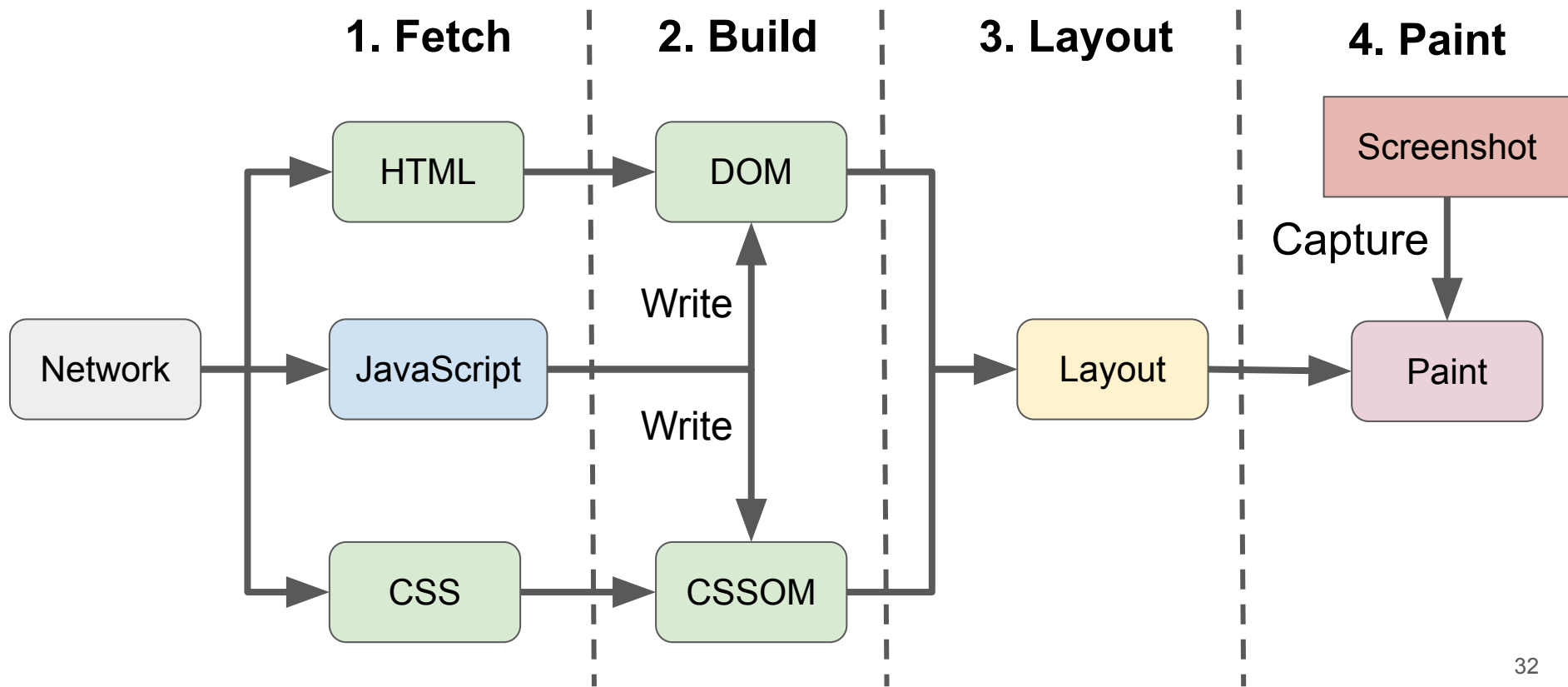
How a browser renders pages



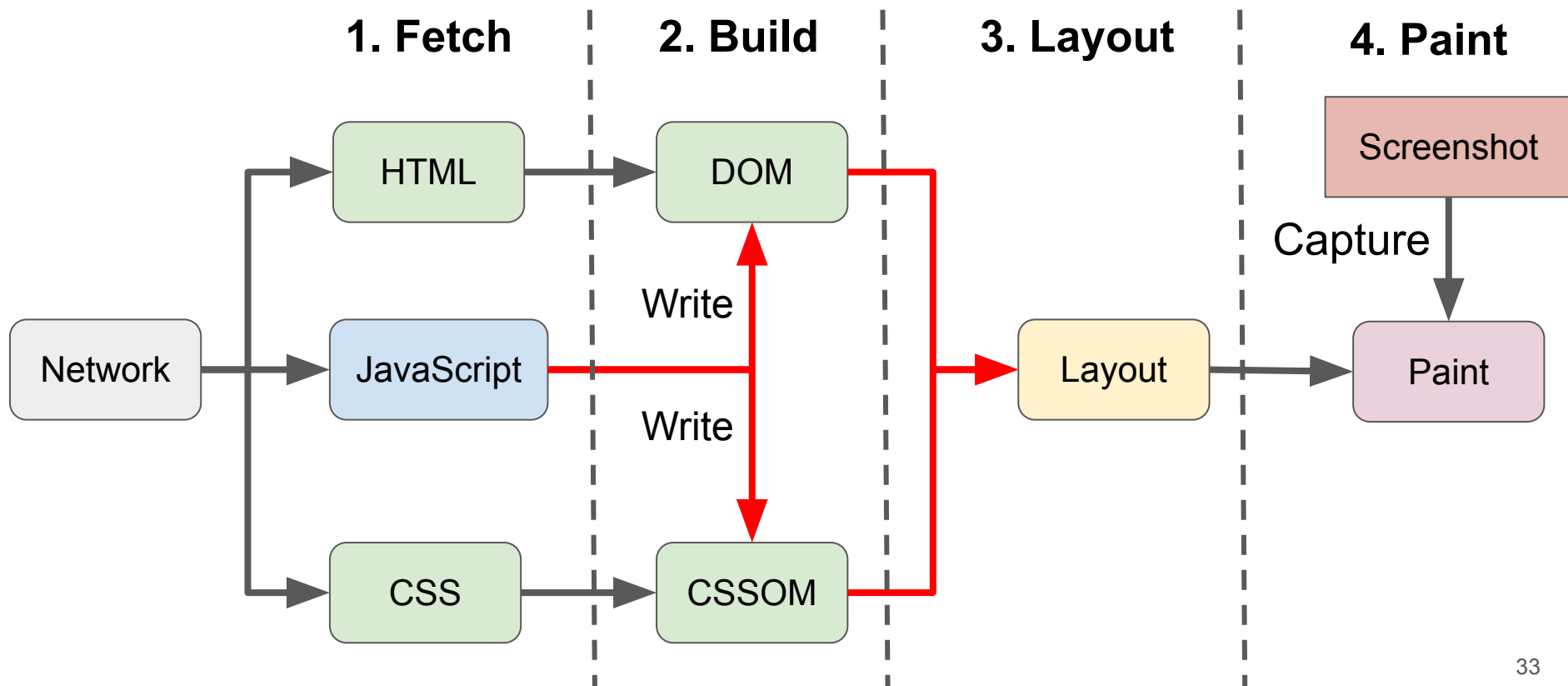
How a browser renders pages



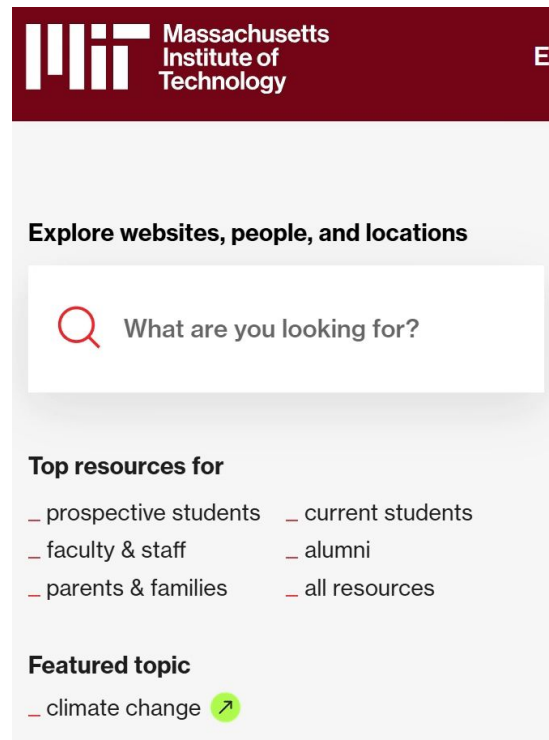
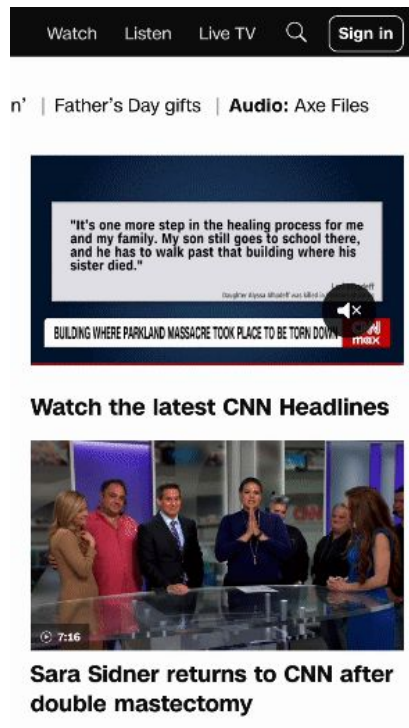
How a browser renders pages



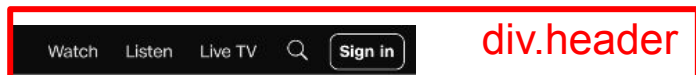
How a browser renders pages



Fidelity violation detection: Match layout tree



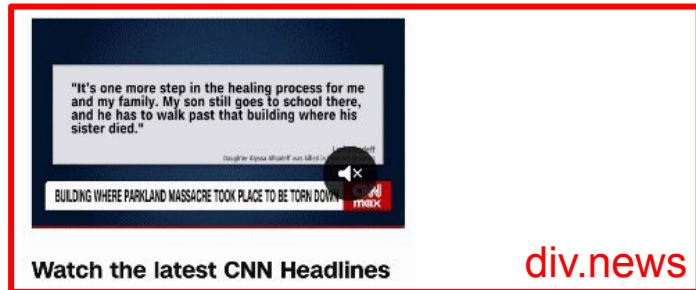
Fidelity violation detection: Match layout tree



div.header



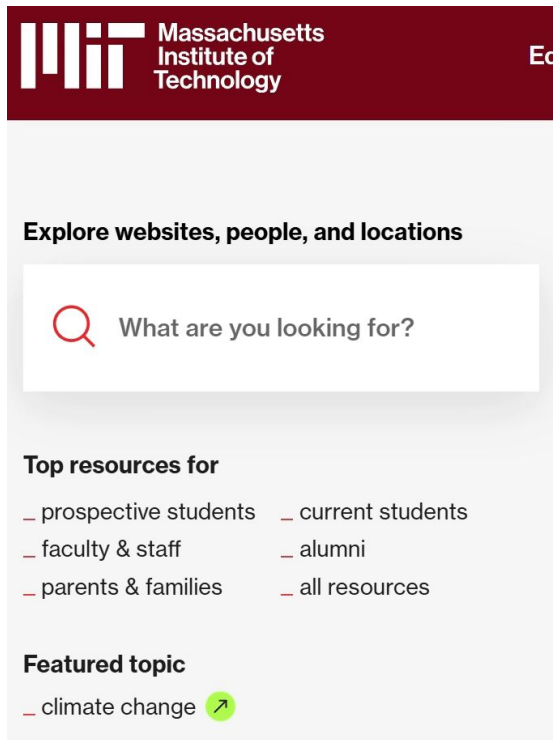
div.links



div.news



div.news



Fidelity violation detection: Match layout tree

Watch Listen Live TV  [Sign in](#)

n' | Father's Day gifts | **Audio:** Axe Files



video

Watch the latest CNN Headlines


div.news




img

Sara Sidner returns to CNN after double mastectomy

div.news

 **Massachusetts
Institute of
Technology** Ed

Explore websites, people, and locations

 What are you looking for?

Top resources for

[_ prospective students](#)

[_ current students](#)


[_ faculty & staff](#)

[_ alumni](#)

[_ parents & families](#)

[_ all resources](#)

Featured topic

[_ climate change](#) 

Fidelity violation detection: Match layout tree

Watch Listen Live TV  [Sign in](#)

n' | Father's Day gifts | **Audio:** Axe Files



video

Watch the latest CNN Headlines


div.news




img

Sara Sidner returns to CNN after double mastectomy

div.news

 **Massachusetts
Institute of
Technology** Ed

Explore websites, people, and locations

 What are you looking for?

Top resources for

[_ prospective students](#)

[_ current students](#)


[_ faculty & staff](#)

[_ alumni](#)

[_ parents & families](#)

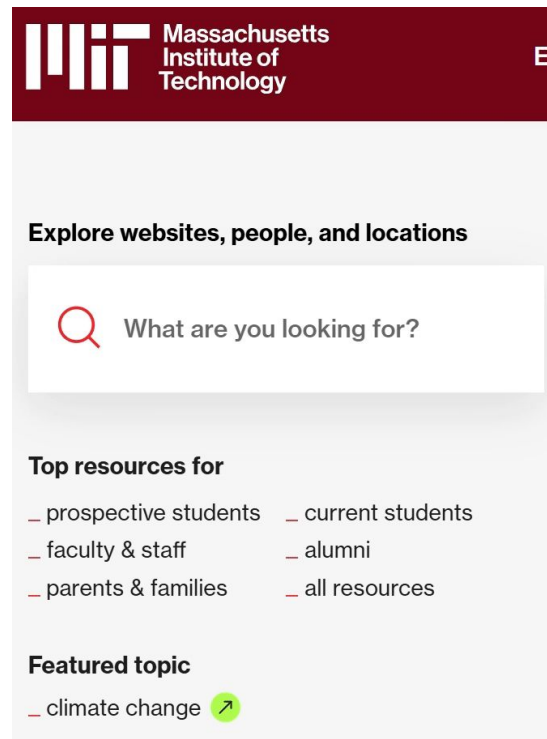
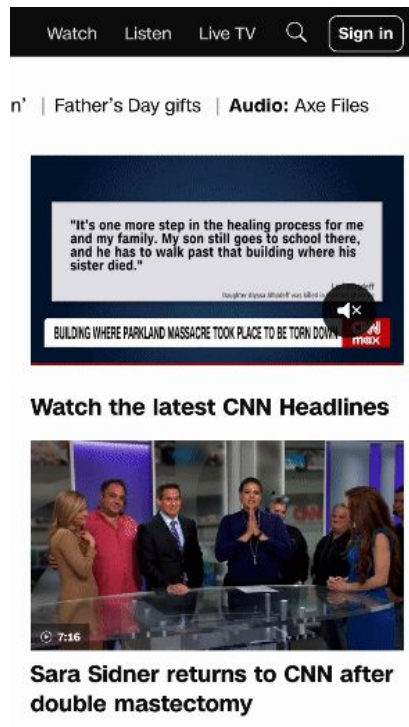
[_ all resources](#)

Featured topic

[_ climate change](#) 

Layout tree unchanged

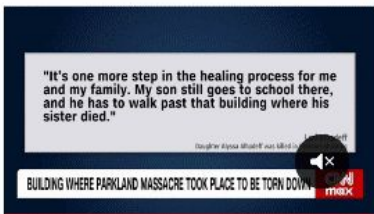
Fidelity violation detection: Match JS writes



Fidelity violation detection: Match JS writes



n' | Father's Day gifts | **Audio:** Axe Files



Watch the latest CNN Headlines



Sara Sidner returns to CNN after double mastectomy



Explore websites, people, and locations

`
{text}
`



What are you looking for?

Top resources for

- _ prospective students
- _ current students
- _ faculty & staff
- _ alumni
- _ parents & families
- _ all resources

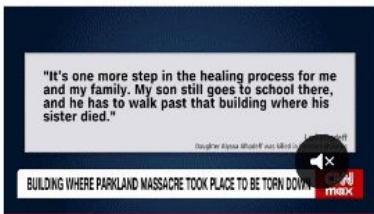
Featured topic

- _ climate change ↗

Fidelity violation detection: Match JS writes



n' | Father's Day gifts | **Audio:** Axe Files



Watch the latest CNN Headlines



Sara Sidner returns to CNN after double mastectomy



Explore websites, people, and locations

**
{text}
**



What are you looking for?

```
for (var n = function(e) {  
  setTimeout(function() {  
    t.inputPlaceholder.textContent = t.searchTerm.substr(0, e)  
  }, t.attractLoopTypeRate * e)  
}, r = 1; r < this.searchTerm.length + 1; r += 1)  
  n(r)
```

Featured topic

climate change ↗

Fidelity violation detection: Match JS writes

[Watch](#) [Listen](#) [Live TV](#) [Sign in](#)

[n'](#) | [Father's Day gifts](#) | **Audio:** [Axe Files](#)


"It's one more step in the healing process for me and my family. My son still goes to school there, and he has to walk past that building where his sister died."

BUILDING WHERE PARKLAND MASSACRE TOOK PLACE TO BE TORN DOWN

Watch the latest CNN Headlines



Sara Sidner returns to CNN after double mastectomy

 **Massachusetts Institute of Technology**

Explore websites, people, and locations

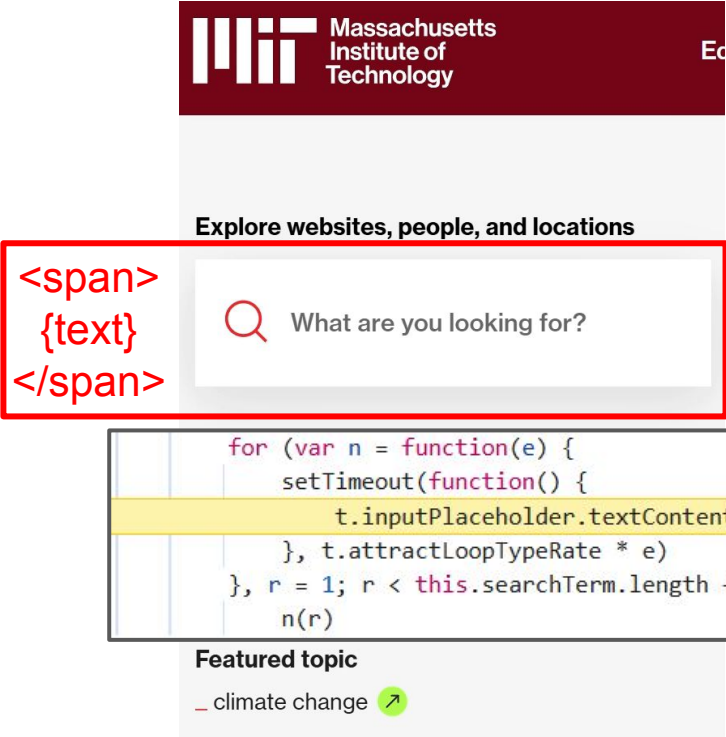
{text}


```
for (var n = function(e) {  
  setTimeout(function() {  
    t.inputPlaceholder.textContent  
  }, t.attractLoopTypeRate * e)  
}, r = 1; r < this.searchTerm.length +  
  n(r)
```

Featured topic
_ climate change ↗

➤ (anonymous)	main.js?siolrj:1
setTimeout	
n	main.js?siolrj:1
value	main.js?siolrj:1
value	main.js?siolrj:1
value	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
setInterval	
(anonymous)	main.js?siolrj:1
setTimeout	
(anonymous)	main.js?siolrj:1
setTimeout	
value	main.js?siolrj:1
t	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
u	main.js?siolrj:1
fireWith	main.js?siolrj:1
r	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
load	
send	main.js?siolrj:1
ajax	main.js?siolrj:1

Fidelity violation detection: Match JS writes



(anonymous)	main.js?siolrj:1
setTimeout	
n	main.js?siolrj:1
value	main.js?siolrj:1
value	main.js?siolrj:1
value	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
setInterval	
(anonymous)	main.js?siolrj:1
setTimeout	
(anonymous)	main.js?siolrj:1
setTimeout	
value	main.js?siolrj:1
t	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
u	main.js?siolrj:1
fireWith	main.js?siolrj:1
r	main.js?siolrj:1
(anonymous)	main.js?siolrj:1
load	
send	main.js?siolrj:1
ajax	main.js?siolrj:1

Associated {JS writes} unchanged

Fidelity violation diagnosis

- Pinpoint causes of violations
 - **Discover** errors not reported by browser
 - Identify if an error **truly matters to fidelity**



Evaluation

- **Datasets**

- 80K pages sampled from [top 1 million sites](#), [5 pages](#) per site
- Pages sampled from End-of-Term and CARTA collections

Evaluation

- **Datasets**

- 80K pages sampled from **top 1 million sites**, **5 pages** per site
- Pages sampled from End-of-Term and CARTA collections

- **Baselines for detecting fidelity violations**

- Match **screenshots**, **HTML extracted texts** (Browsertrix's QA)
- Check if archive **has extra errors**

FidEx improves accuracy of detecting fidelity violations

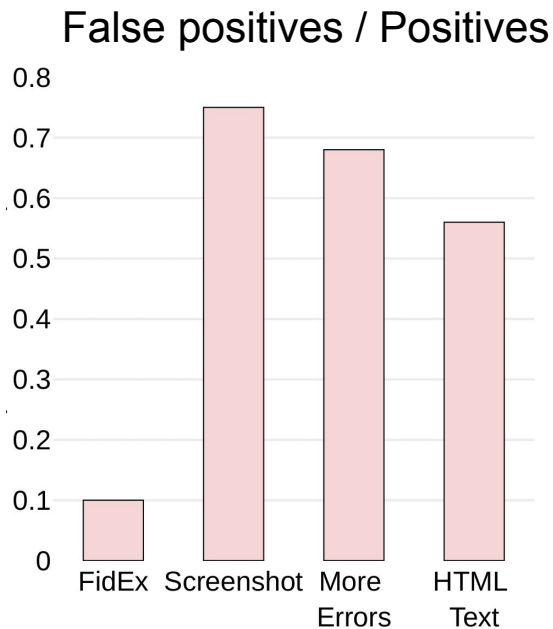
	FidEx	Screenshots	More Errors	HTML Text
Positive rate	15.5%	72.6%	43.7%	24.9%

*Positive:
detected fidelity
violation(s)*

FidEx improves accuracy of detecting fidelity violations

	FidEx	Screenshots	More Errors	HTML Text
Positive rate	15.5%	72.6%	43.7%	24.9%

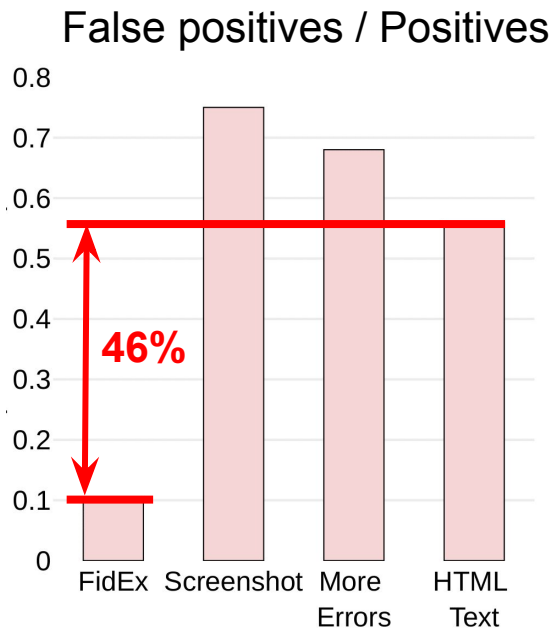
*Positive:
detected fidelity
violation(s)*



FidEx improves accuracy of detecting fidelity violations

	FidEx	Screenshots	More Errors	HTML Text
Positive rate	15.5%	72.6%	43.7%	24.9%

*Positive:
detected fidelity
violation(s)*

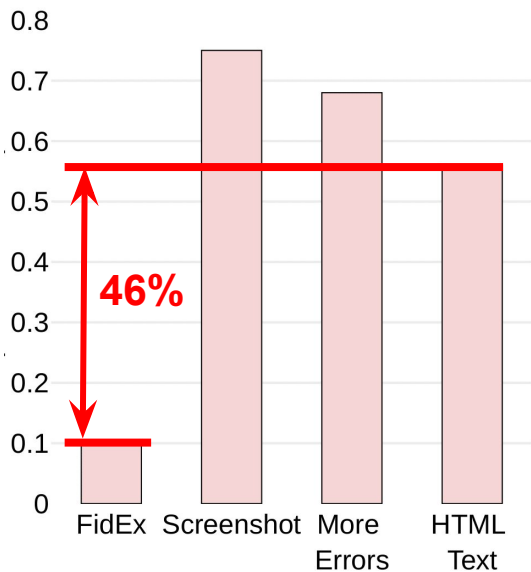


FidEx improves accuracy of detecting fidelity violations

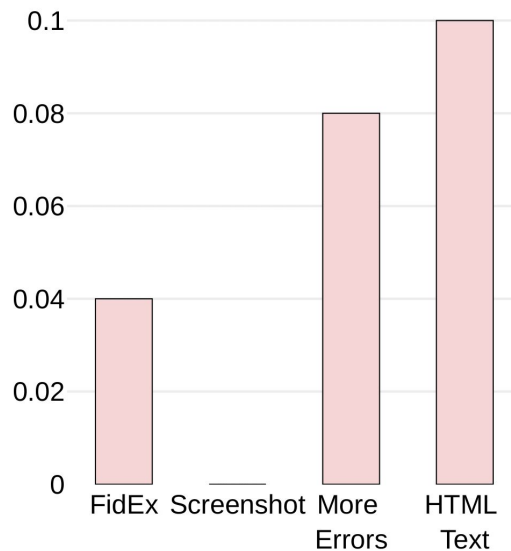
	FidEx	Screenshots	More Errors	HTML Text
Positive rate	15.5%	72.6%	43.7%	24.9%

*Positive:
detected fidelity
violation(s)*

False positives / Positives



False negatives / Negatives

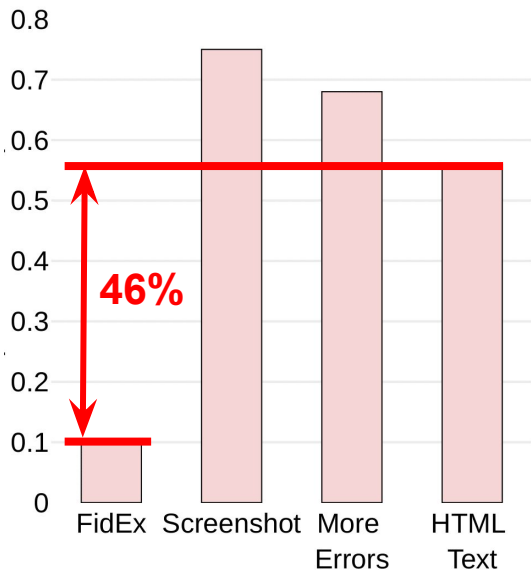


FidEx improves accuracy of detecting fidelity violations

	FidEx	Screenshots	More Errors	HTML Text
Positive rate	15.5%	72.6%	43.7%	24.9%

*Positive:
detected fidelity
violation(s)*

False positives / Positives



False negatives / Negatives



**If classify every
page as
negative, false
negative rate
will be ~15%**

Fidelity Violations Detected by FidEx in EOT

Live web page

Chapter 13 - Bankruptcy Basics

This chapter of the Bankruptcy Code provides for adjustment of debts of an individual with regular income. Chapter 13 allows a debtor to keep property and pay debts over time, usually three to five years.

Background

A chapter 13 bankruptcy is also called a wage earner's plan. It enables individuals with regular income to develop a plan to repay all or part of their debts. Under this chapter, debtors propose a repayment plan to make installments to creditors over three to five years. If the debtor's current monthly income is less than the applicable state median, the plan will be for three years unless the court approves a longer period "for cause." (1) If the debtor's current monthly income is greater than the applicable state median, the plan generally must be for five years. In no case may a plan provide for payments over a period longer than five years. 11 U.S.C. § 1322(d). During this time the law forbids creditors from starting or continuing collection efforts.

This chapter discusses six aspects of a chapter 13 proceeding: the advantages of choosing chapter 13, the chapter 13 eligibility requirements, how a chapter 13 proceeding works, making the plan work, and the special chapter 13 discharge.

Federal Court Finder

Location Court Name

Archived copy

Chapter 13 - Bankruptcy Basics

This chapter of the Bankruptcy Code provides for adjustment of debts of an individual with regular income. Chapter 13 allows a debtor to keep property and pay debts over time, usually three to five years.

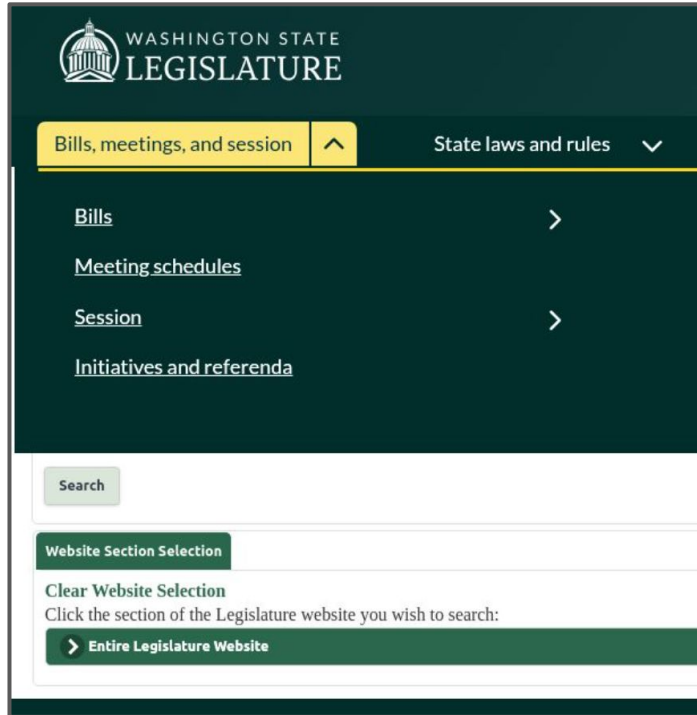
Background

A chapter 13 bankruptcy is also called a wage earner's plan. It enables individuals with regular income to develop a plan to repay all or part of their debts. Under this chapter, debtors propose a repayment plan to make installments to creditors over three to five years. If the debtor's current monthly income is less than the applicable state median, the plan will be for three years unless the court approves a longer period "for cause." (1) If the debtor's current monthly income is greater than the applicable state median, the plan generally must be for five years. In no case may a plan provide for payments over a period longer than five years. 11 U.S.C. § 1322(d). During this time the law forbids creditors from starting or continuing collection efforts.

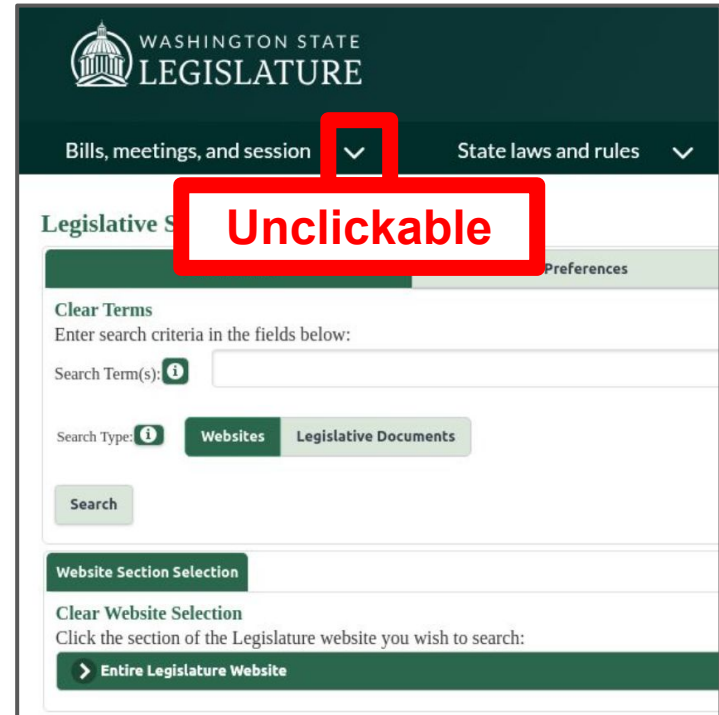
This chapter discusses six aspects of a chapter 13 proceeding: the advantages of choosing chapter 13, the chapter 13 eligibility requirements, how a chapter 13 proceeding works, making the plan work, and the special chapter 13 discharge.

Fidelity Violations Detected by FidEx in EOT

Live web page



Archived copy

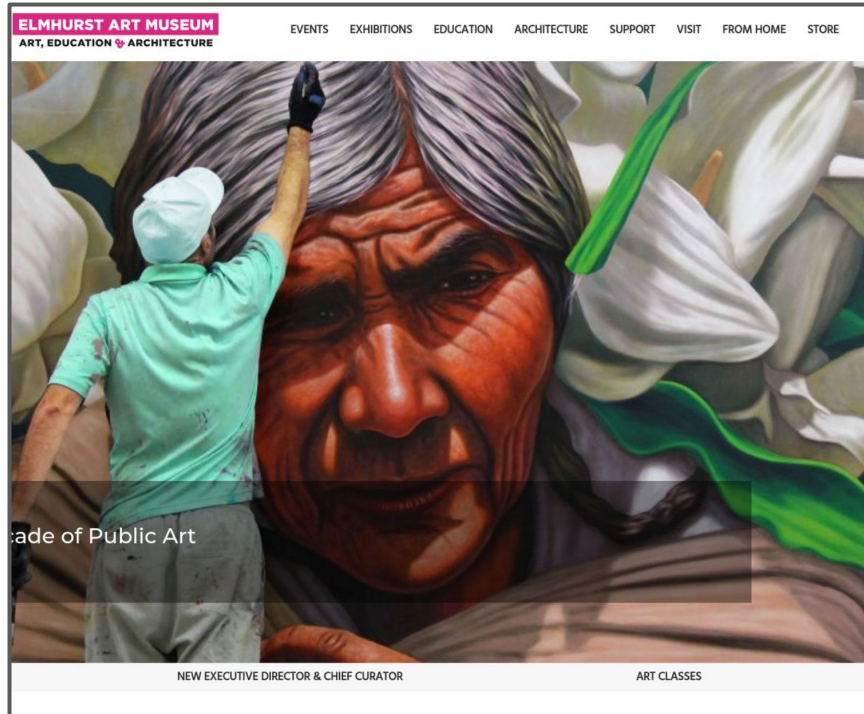


False Positives with Screenshot Comparison in EOT

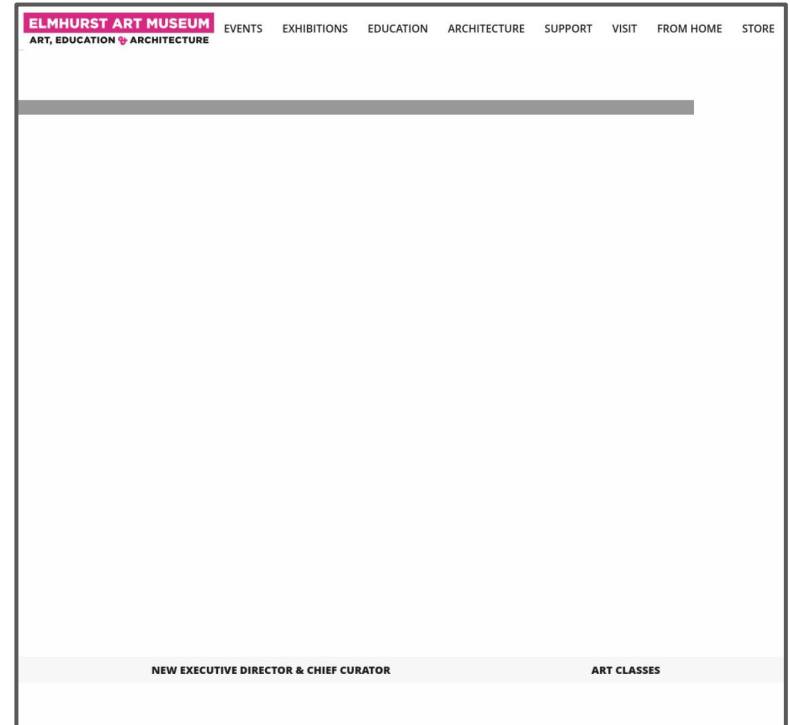


Fidelity Violations Detected by FidEx in CARTA

Live web page

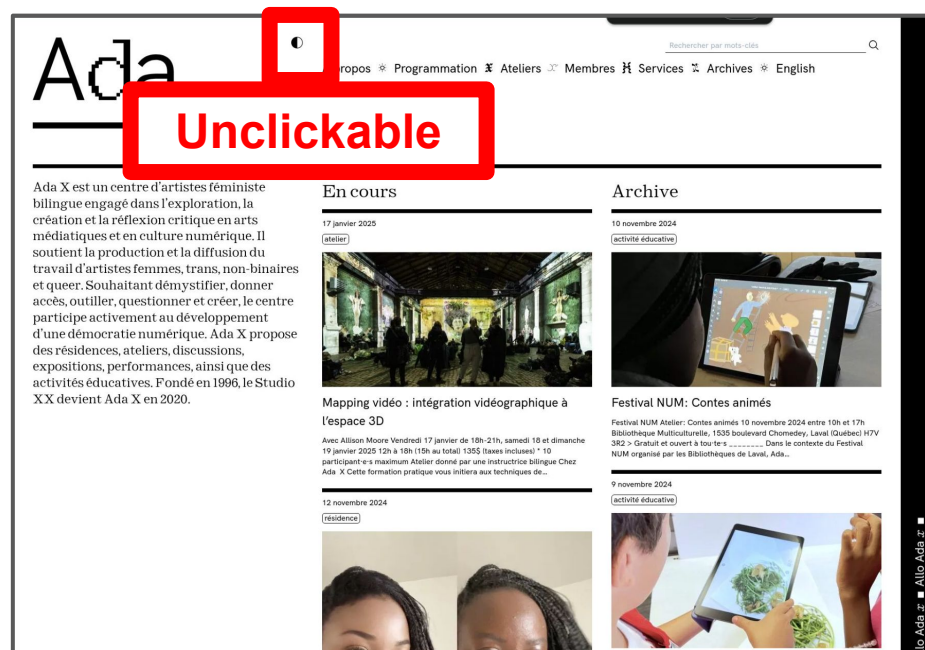


Archived copy

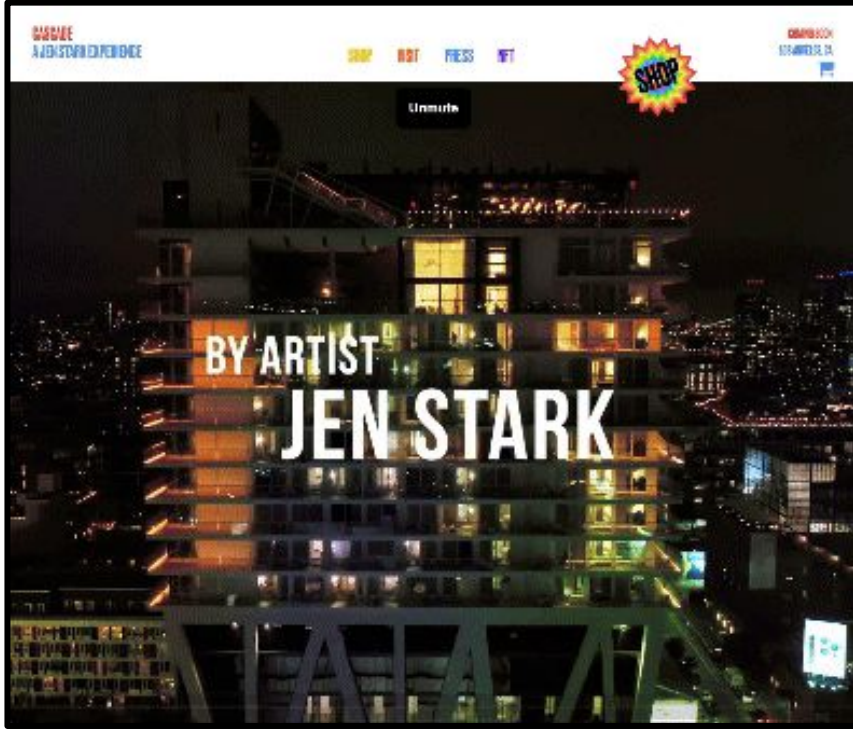


Live web page

Archived copy



False Positives with Screenshot Comparison in CARTA



Summary

- Bugs in JS rewrites → **Poor fidelity replay** of archived pages

Summary

- Bugs in JS rewrites → **Poor fidelity replay** of archived pages

Detection

- **Screenshot comparison**
leads to false positives
- Instead, **compare layout**
tree and **JS writes**

Summary

- Bugs in JS rewrites → **Poor fidelity replay** of archived pages

Detection

- **Screenshot comparison leads to false positives**
- Instead, **compare layout tree** and JS writes

Diagnosis

- **Discover errors not reported** by browser
- Confirm **which errors are worth investigating**

Thank you!
Q & A