CSE227 – Graduate Computer Security

Web Fundamentals / Web Privacy

UC San Diego

Housekeeping

General course things to know

- *Nothing on my mind!*

Today's lecture

Learning Objectives

- Discuss the HTML sanitization paper
- Discuss the two web tracking papers

Parse Me Baby One More Time: Bypassing HTML Sanitizer via Parsing Differentials

What is HTML?

What is HTML?

Hypertext Markup Language: The **structure** of how we embed web content into web pages.

What is Cross-Site Scripting (XSS)?

What is Cross-Site Scripting (XSS)?

"Cross-Site Scripting (XSS) attacks are a type of injection, in which malicious scripts are injected into otherwise benign and trusted websites" – OWASP

Very simple XSS example

Web Server

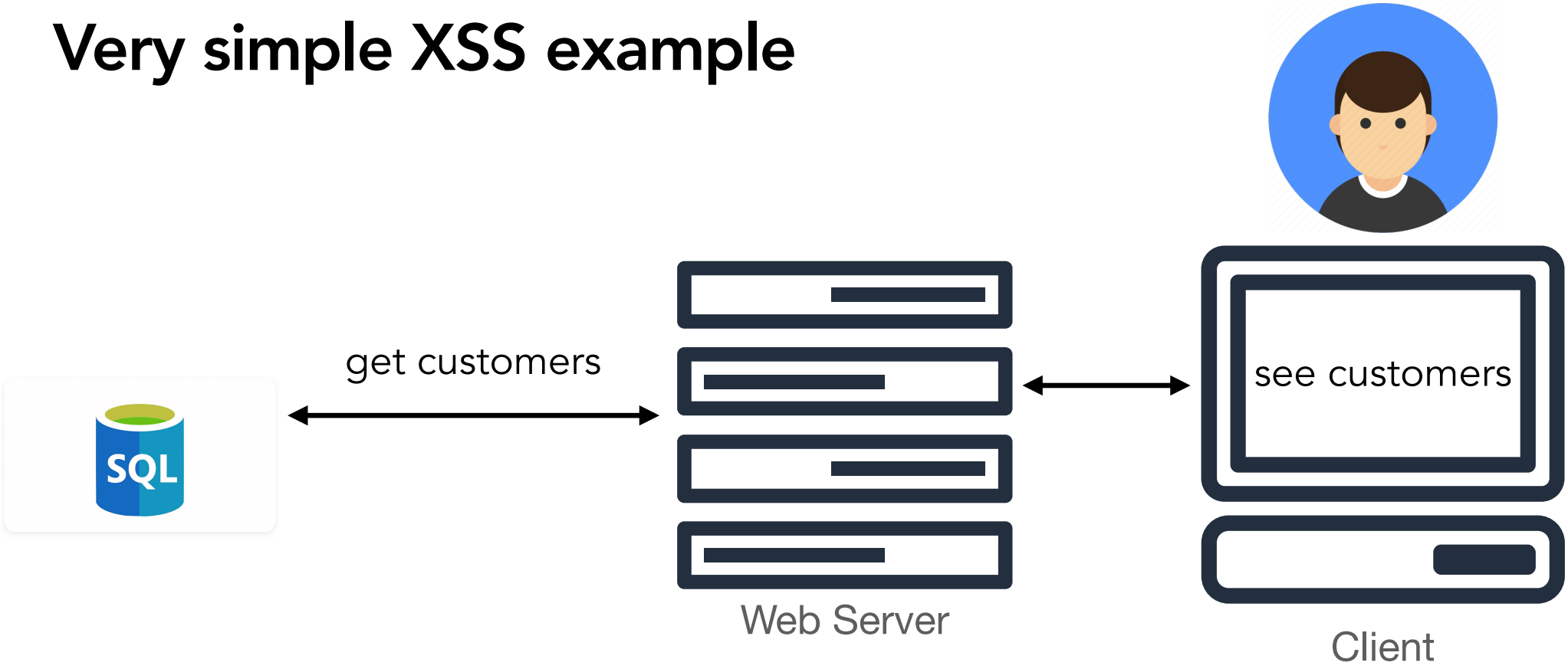
Client

Very simple XSS example

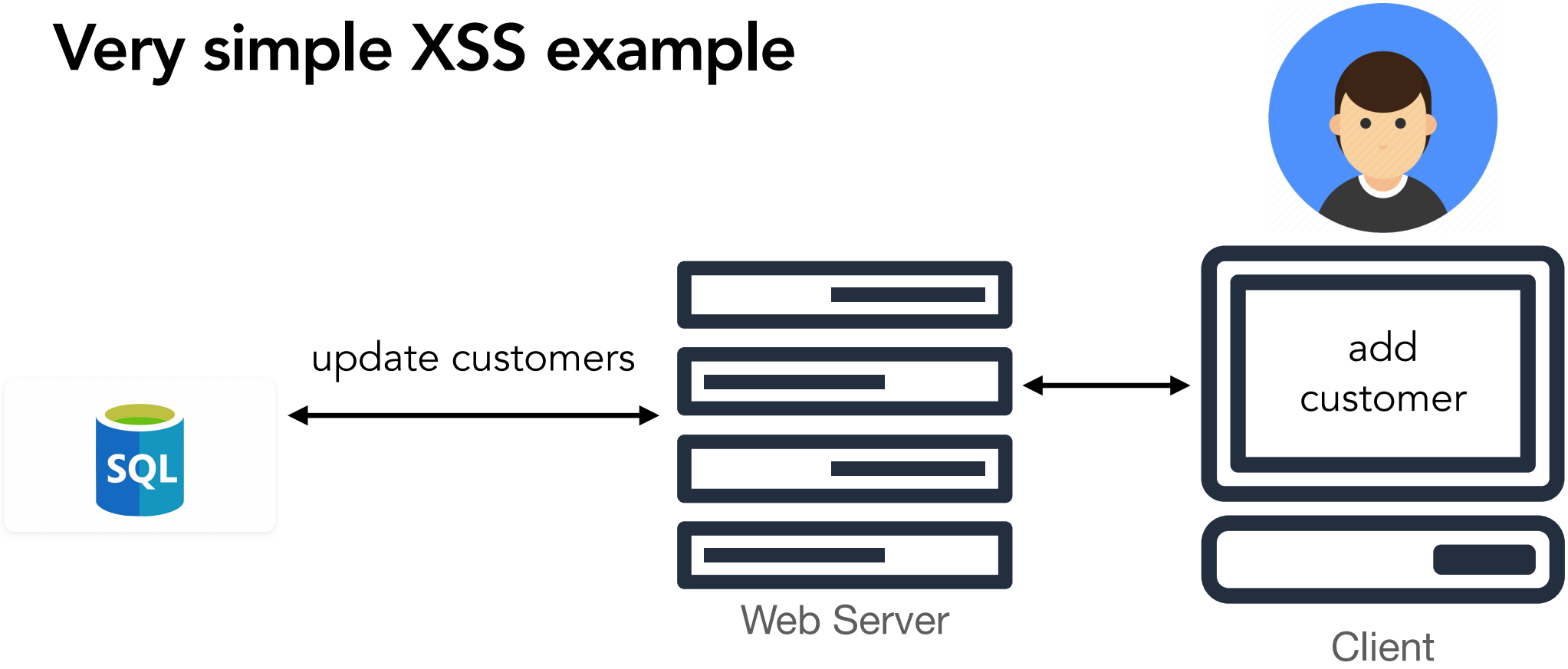
Web Server

Client

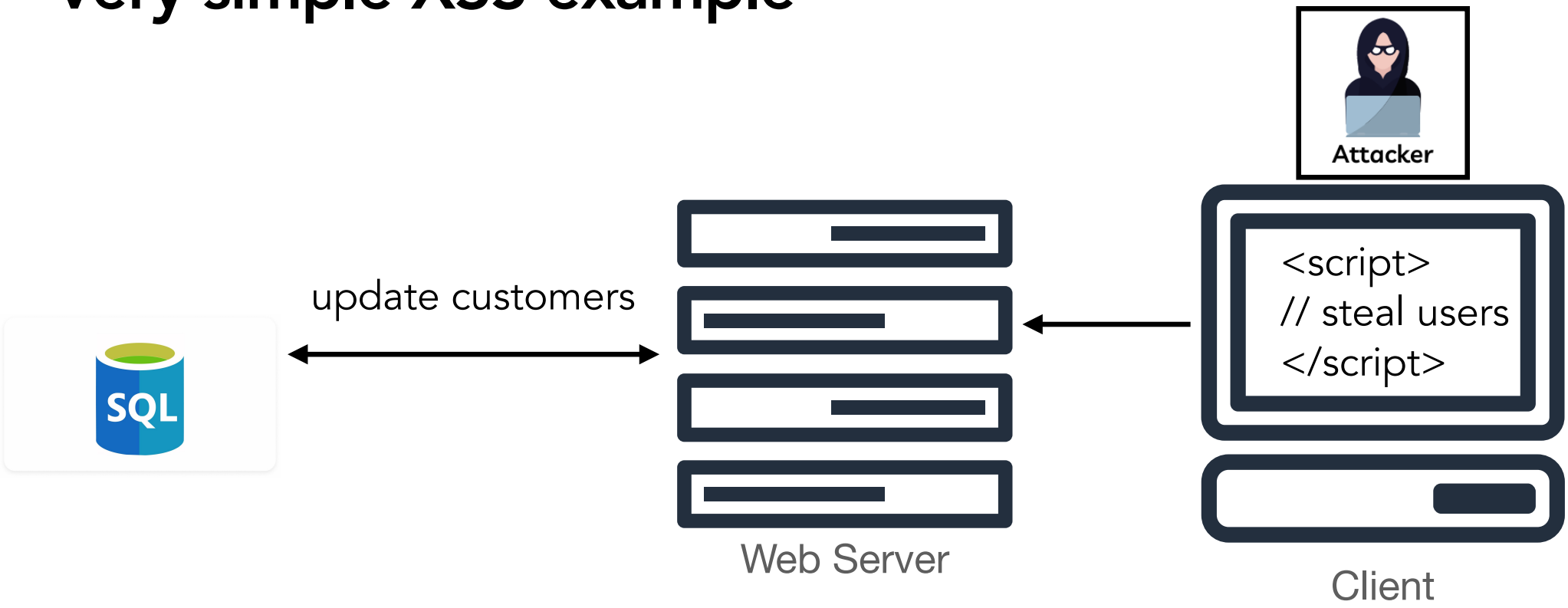
Very simple XSS example



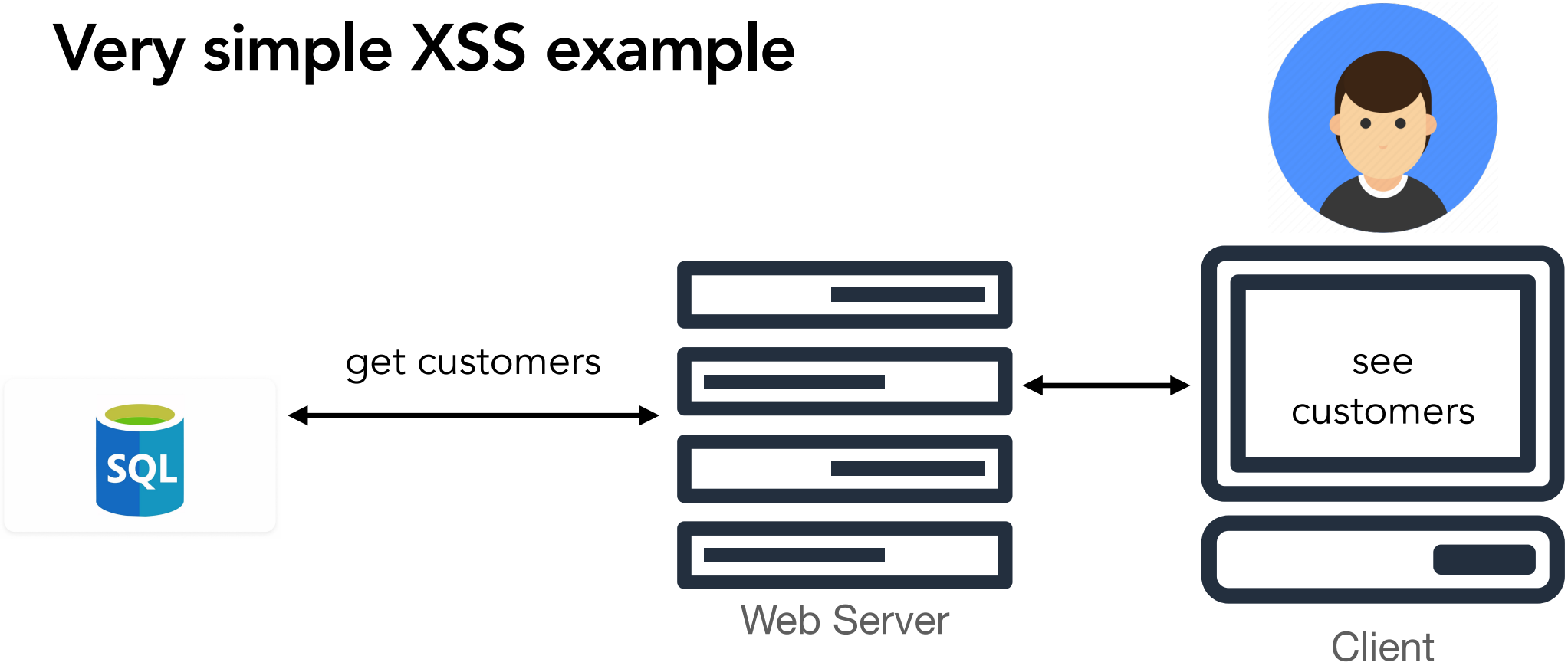
Very simple XSS example



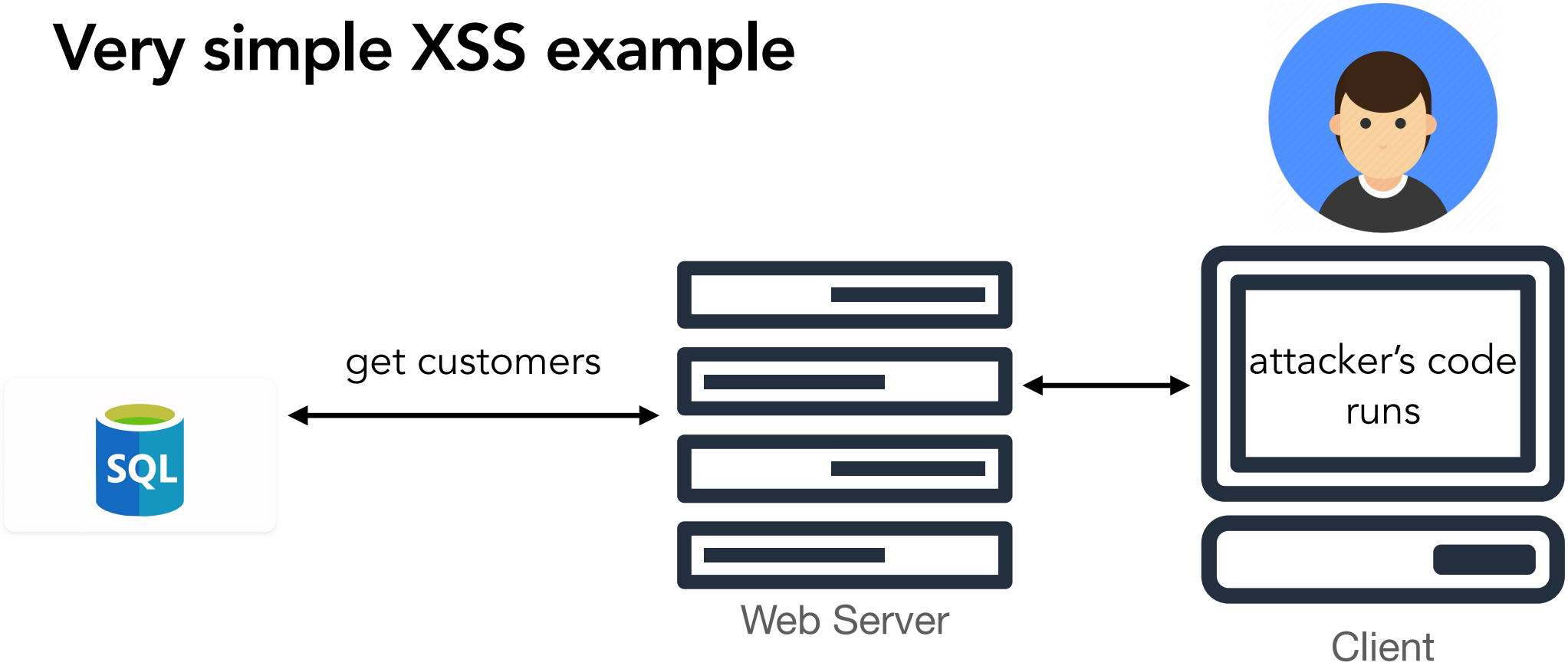
Very simple XSS example



Very simple XSS example



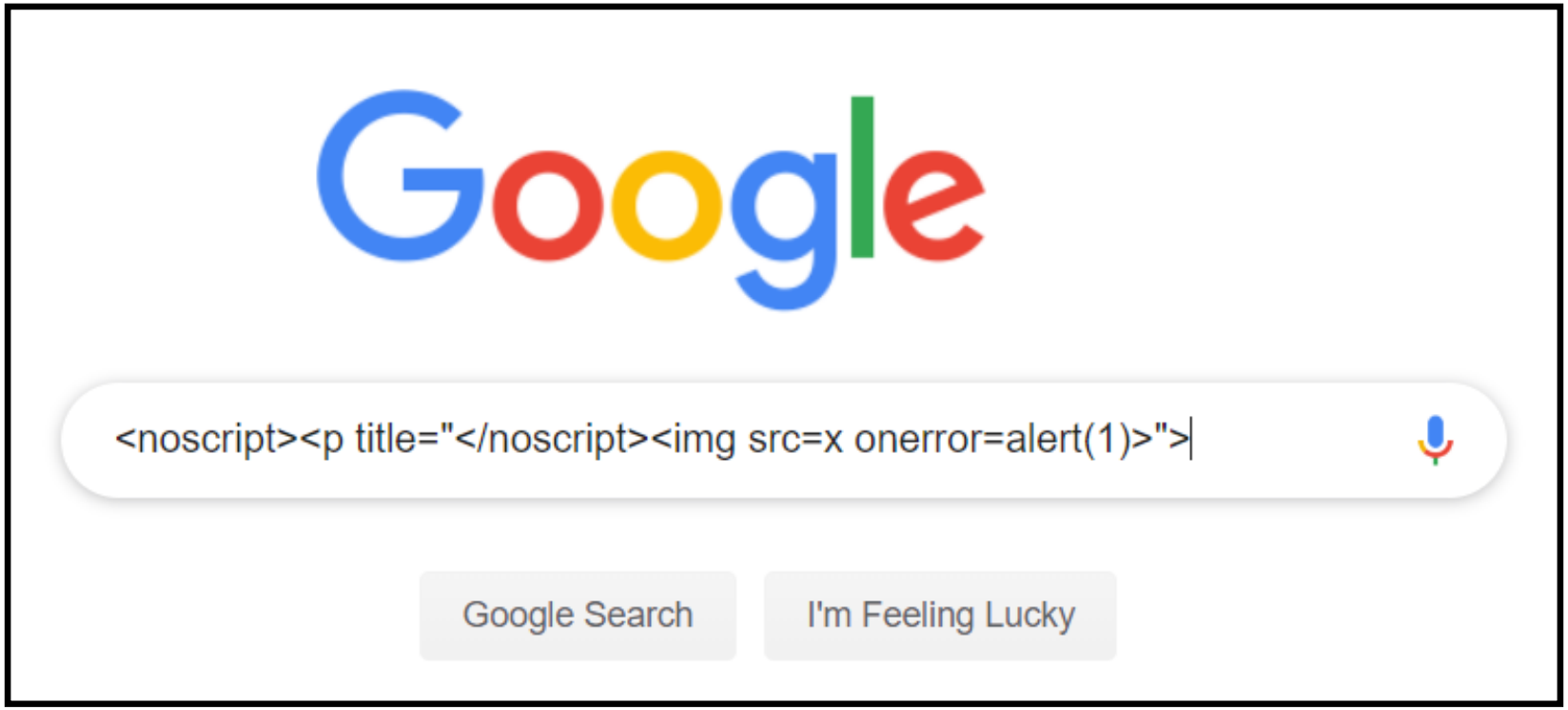
Very simple XSS example



What is Mutation Cross-Site Scripting (mXSS)?

"Such a vulnerability occurs if an HTML fragment is parsed, serialized, and yields a different result upon being parsed again."

mXSS example: Google Search in 2019



Common XSS defenses

- How do we defend against XSS attacks?

Common XSS defenses

- How do we defend against XSS attacks?
- What is input sanitization?

Common XSS defenses

- How do we defend against XSS attacks?
- What is input sanitization?
- Where does input sanitization happen? On the client side or server side?

Issues with server-side sanitization

- Why is accurate HTML sanitization quite hard for servers to do?

Issues with server-side sanitization

- Why is accurate HTML sanitization quite hard for servers to do?
  - Context dependent
  - Requires understanding how the browser is going to interpret the HTML, which turns out is not easy!

This paper asks two questions

1. Is server-side sanitization even feasible (does not ruin benign content) and is secure?

2. How do popular open-source libraries fare in parsing and sanitizing HTML content for XSS attacks?

Their setup for evaluating parsing differentials

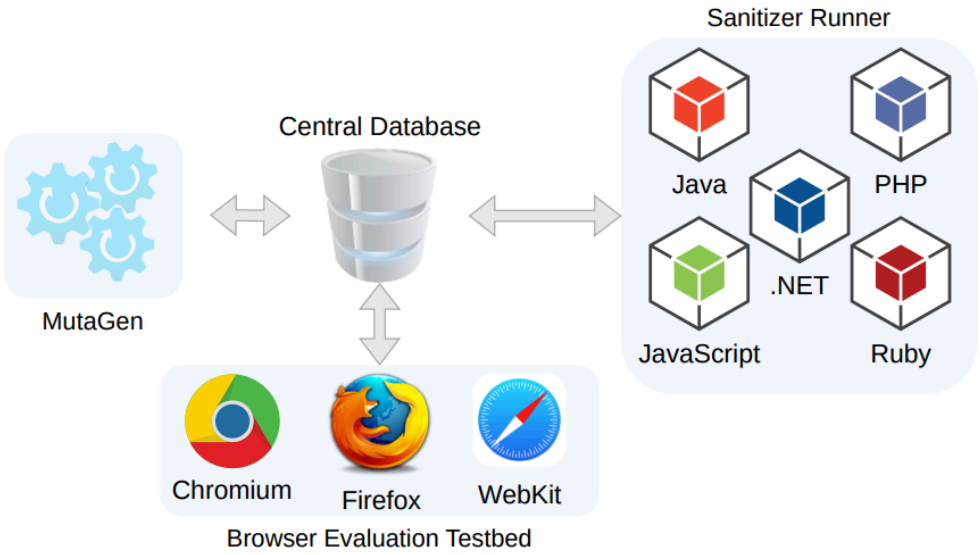


Figure 3: Sanitizer Evaluation Setup

Mutagen: Generating HTML Fragments

- Not going to get into the details here (there are many fine points) – general gist is as follows:
  - Start with a payload *P* that you're sure works
  - Make some transformations to *P* you think might be tricky for a browser
  - Test your set of transformations and keep the ones you think work
  - Repeat with new P

Parsing differential strategy

- Tested 11 (really 12) different parsers (common libraries), throwing all the generated mutations into each one, and saw how they compared
- How did the authors evaluate if the parsing was as they expected it to be?

Results

- What did the authors find as their top-line results?
- Did every browser interpret HTML identically? Which browsers didn't?
- What do these results tell us about HTML parsing?

Sanitizer	Chrome		Webkit		Firefox	
	F	D	F	D	F	D
DOMPurify	0.87	0.87	0.87	0.87	0.81	0.86
DOMPurify (jsdom19)	0.88	0.88	0.88	0.88	0.82	0.88
sanitizer	0.36	0.36	0.36	0.36	0.37	0.36
google-caja-sanitizer	0.50	0.50	0.50	0.50	0.50	0.50
sanitize-html	0.39	0.39	0.39	0.39	0.41	0.39
HtmlSanitizer	0.90	0.90	0.90	0.90	0.84	0.90
HtmlRuleSanitizer	0.15	0.15	0.15	0.15	0.15	0.15
Typo3	0.52	0.52	0.52	0.52	0.53	0.52
rgrove/sanitize	0.94	0.94	0.94	0.94	0.88	0.94
loofah	0.22	0.22	0.22	0.22	0.25	0.22
AntiSamy	0.58	0.58	0.58	0.58	0.58	0.58
JSoup	0.51	0.51	0.51	0.51	0.52	0.51

F: fragment parsing, D: document parsing

Combining our two papers…

- How does XSS related to CSRF?
- Do CSRF defenses protect against XSS?
- What is the relationship between XSS and CSRF?
- What would you say is a **fundamental issue** that enables a XSS attack?

Combining our two attacks…

- How does XSS related to CSRF?
- Do CSRF defenses protect against XSS?
- What is the relationship between XSS and CSRF?
- What would you say is a **fundamental issue** that enables a XSS attack?
  - Mixing code and data!

Paper meta-questions

- What did we think about the paper?
  - You can comment on the organization, the writing, the experiments, etc.
- What do you think about the solution presented in the paper?
- Why do you think this paper was so successful?

Break Time + Attendance

https://tinyurl.com/cse227-attend

Codeword:
Excesses

The Web Never Forgets: Persistent Tracking Mechanisms in the Wild

What is web tracking?

What is web tracking?

A suite of technologies designed to collect, analyze, and track user activity on the web.

Why does web tracking exist?

Why does web tracking exist?

- Advertising – people make dollars off of targeting **you** on the web
  - The more targeted your advertising, the more revenue you can make from advertisers who are potentially willing to give you more money to sell the ad spot
  - Useful for advertisers to know if people with your browsing habits, your properties, your whatever are browsing on the web

Trump escalates January 6 cover-up

The former President is trying to keep the House select committee probing January 6 from seeing a list of documents as he ramps up his political comeback

KFILE Trump lawyer said 'courage and the spine' would help Pence send election to the House in comments before January 6

Brian Stelter's ominous prediction: Imagine it's 2022 and ...

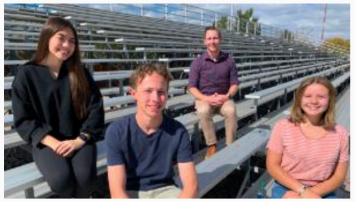
January 6 committee is losing patience with Trump's former chief of staff Mark Meadows as it seeks his testimony

Washington Post report rebuts the January 6 alt-reality that Tucker Carlson promotes

Biden says US 'continuing to suffer' from Trump's decision to pull out of Iran nuclear deal

Astros top Braves 9-5 in World Series Game 5

• **Trivia:** Can you name the only player to play in all 3 cities that the Braves have called home?
• **Analysis:** The Braves may win the World Series. But they're striking out with some fans



Students are fed up with raging adults at school board meetings

• A Texas lawmaker is investigating 850 books on race and gender that could cause 'discomfort' to students
• **Opinion:** When parents scream at school board meetings, how can I teach their children?

Southwest launches investigation into pilot reportedly using anti-Biden phrase on flight

Reporter reveals what Lindsey Graham said during January 6 riot

White House press secretary tests positive for Covid, last saw Biden Tuesday

BREAKING Japan's Fumio Kishida defies expectations as ruling party keeps majority

Aurora borealis puts on a gorgeous show

'Step up or step out': Lawmaker calls out attorney general

Police investigating desecration of Torah scroll at fraternity

COP26 climate talks off to an ominous start after weak G20 leaders' meeting

Video shows passengers fleeing knife attack on train




Trump escalates January 6 cover-up

The former President is trying to keep the House select committee probing January 6 from seeing a list of documents as he ramps up his political comeback

KFILE Trump lawyer said 'courage and the spine' would help Pence send election to the House in comments before January 6

Brian Stelter's ominous prediction: Imagine it's 2022 and ...

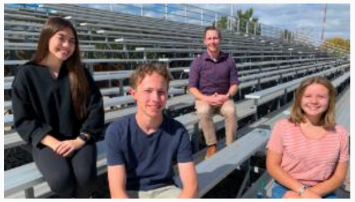
January 6 committee is losing patience with Trump's former chief of staff Mark Meadows as it seeks his testimony

Washington Post report rebuts the January 6 alt-reality that Tucker Carlson promotes

Biden says US 'continuing to suffer' from Trump's decision to pull out of Iran nuclear deal

Astros top Braves 9-5 in World Series Game 5

• **Trivia:** Can you name the only player to play in all 3 cities that the Braves have called home?
• **Analysis:** The Braves may win the World Series. But they're striking out with some fans



Students are fed up with raging adults at school board meetings

• A Texas lawmaker is investigating 850 books on race and gender that could cause 'discomfort' to students
• **Opinion:** When parents scream at school board meetings, how can I teach their children?

Southwest launches investigation into pilot reportedly using anti-Biden phrase on flight

Reporter reveals what Lindsey Graham said during January 6 riot

White House press secretary tests positive for Covid, last saw Biden Tuesday

BREAKING Japan's Fumio Kishida defies expectations as ruling party keeps majority

Aurora borealis puts on a gorgeous show

'Step up or step out': Lawmaker calls out attorney general

Police investigating desecration of Torah scroll at fraternity

COP26 climate talks off to an ominous start after weak G20 leaders' meeting

Video shows passengers fleeing knife attack on train




Web Tracking

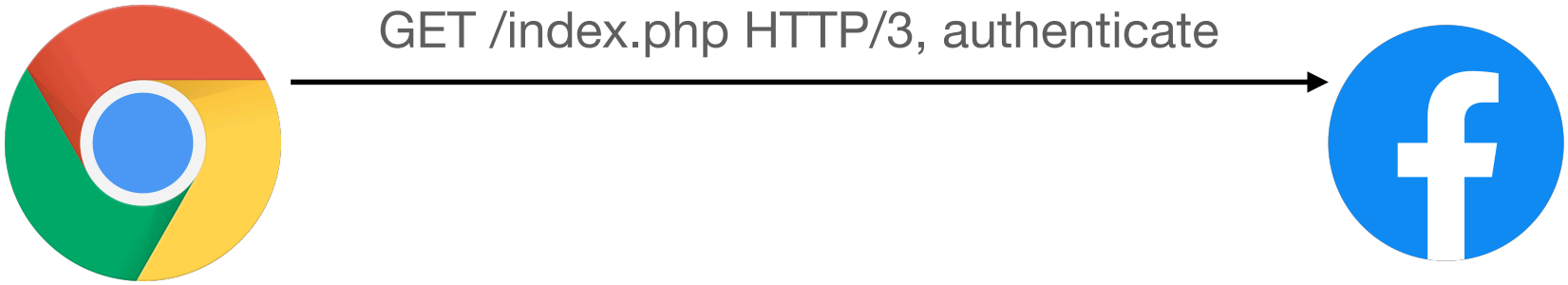
Cookies and Code

- Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)

Web Tracking

Cookies and Code

- Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)



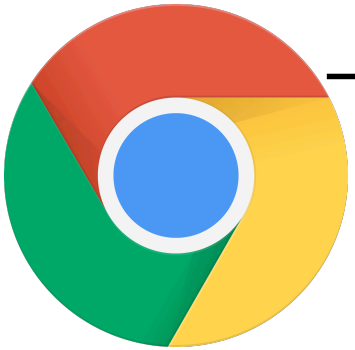
Web Tracking

Cookies and Code

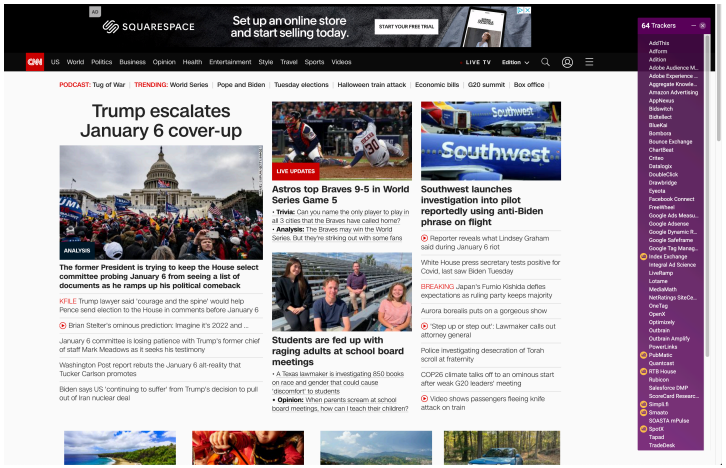
- Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)

Web Tracking

Cookies and Code

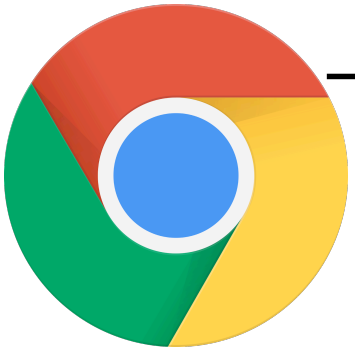


GET / HTTP/3



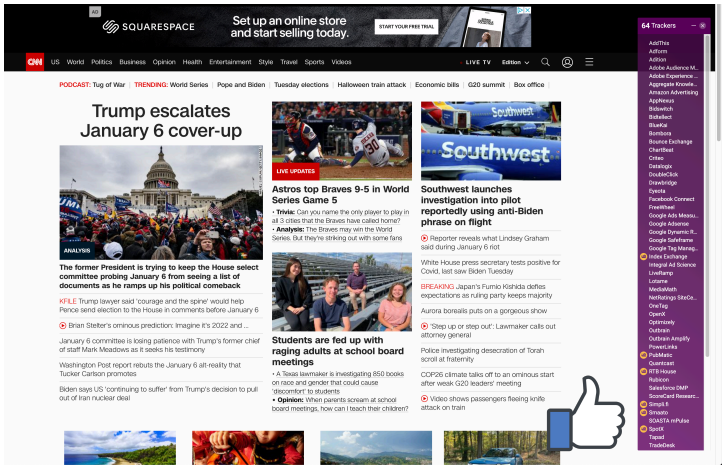
Web Tracking

Cookies and Code



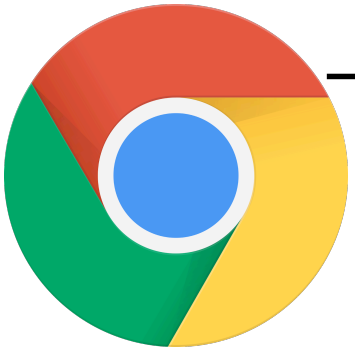
GET / HTTP/3

GET /facebook-like.js HTTP/3



Web Tracking

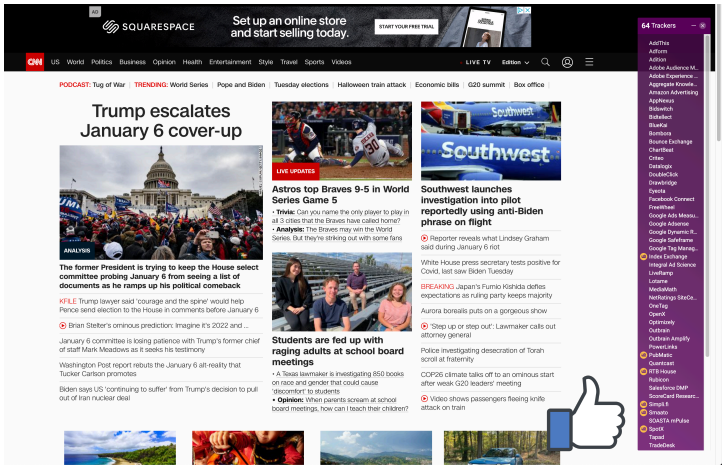
Cookies and Code



GET / HTTP/3

GET /facebook-like.js HTTP/3

Cookie: User=Deepak, Referer=cnn.com



- With this request, companies can link your cookie to your browsing data (e.g., through Referer header, Host headers, Origin, or just JavaScript)

This paper: Three mechanisms in depth

- Canvas Fingerprinting
- Evercookies
- Cookie syncing

Canvas Fingerprinting

- What is canvas fingerprinting?

Canvas Fingerprinting

- What is canvas fingerprinting?
  - Using the Canvas API, adversary can exploit subtle differences in the rendering of the same text or drawing to extract a consistent fingerprint

Canvas Fingerprinting

- What is canvas fingerprinting?
  - Using the Canvas API, adversary can exploit subtle differences in the rendering of the same text or drawing to extract a consistent fingerprint
- Why are fingerprints so unique?

Canvas Fingerprinting

- What is canvas fingerprinting?
  - Using the Canvas API, adversary can exploit subtle differences in the rendering of the same text or drawing to extract a consistent fingerprint
- Why are fingerprints so unique?
  - Many factors go into fingerprints: browser version, installed fonts, graphics cards, graphics drivers, etc.

Canvas Fingerprinting

- What is canvas fingerprinting?
  - Using the Canvas API, adversary can exploit subtle differences in the rendering of the same text or drawing to extract a consistent fingerprint
- Why are fingerprints so unique?
  - Many factors go into fingerprints: browser version, installed fonts, graphics cards, graphics drivers, etc.
- amiunique.org – my computer is unique among 3.3M fingerprints tested

Measuring Canvas Fingerprinting

- At a high level, what was the strategy the authors used to identify canvas fingerprinting in the wild?

Measuring Canvas Fingerprinting

- At a high level, what was the strategy the authors used to identify canvas fingerprinting in the wild?
- What % of the top 100K websites had canvas fingerprinting enabled on the home page?

Measuring Canvas Fingerprinting

- At a high level, what was the strategy the authors used to identify canvas fingerprinting in the wild?
- What % of the top 100K websites had canvas fingerprinting enabled on the home page?
  - 5542 (approximately 5% of websites) – most of which were third party ad companies

Measuring Canvas Fingerprinting

- At a high level, what was the strategy the authors used to identify canvas fingerprinting in the wild?
- What % of the top 100K websites had canvas fingerprinting enabled on the home page?
  - 5542 (approximately 5% of websites) – most of which were third party ad companies
  - Is 5% bad? Good?

Evercookies

- What are evercookies?

Evercookies

- What are evercookies?
  - Resilient tracking mechanism that utilizes multiple storage vectors: localStorage, sessionStorage, Flash, ETags, service workers, chrome extensions, etc.
- Basic idea is: if you store it in multiple places with access by **origin**, then it can be replaced if you can get JavaScript running on the browser (very easy)

Measuring Evercookies

- How did the authors measure evercookies?

Measuring Evercookies

- How did the authors measure evercookies?
  - Crawl 1: 10000 sites, save all cookies
  - Crawl 2, 3: 10000 sites, loaded with Flash cookies
  - Crawl 4: 10000 sites, no data from previous crawl
- Essentially, selectively remove cookies and see if they get “respawned,” and by whom
  - Authors found many third-parties participating in this practice

Cookie Syncing

- What is cookie syncing?

Cookie Syncing

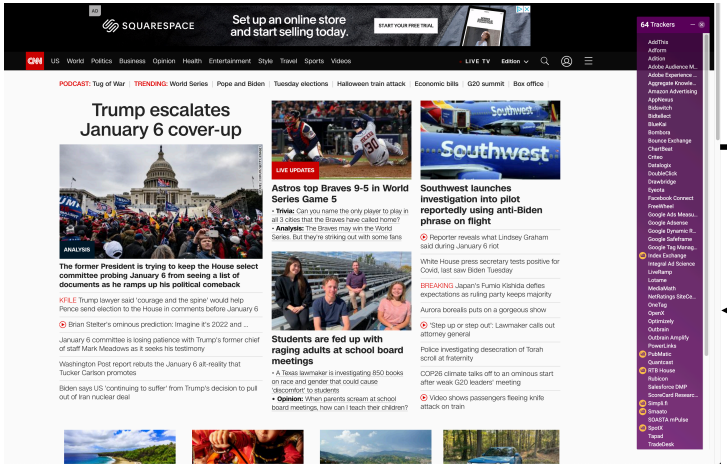
- What is cookie syncing?
  - Tracker domains passing pseudonyms associated with a given user by making **requests** to that website with values as parameters
- How does this practice circumvent the same-origin policy?

Cookie Syncing

- What is cookie syncing?
  - Tracker domains passing pseudonyms associated with a given user by making **requests** to that website with values as parameters
- How does this practice circumvent the same-origin policy?
  - Cookies can't be read by different origins, but if you tell them directly…. then it's all "above board"

Web Tracking

Cookie Syncing



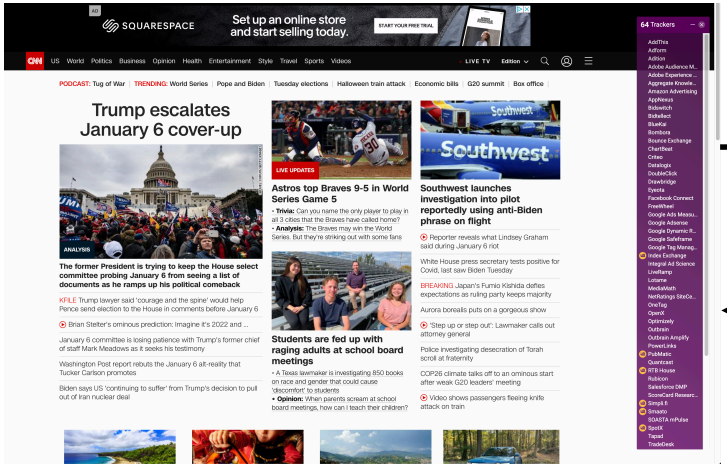
GET tracker.com/pixel.jpg

Response, Set-Cookie: User=user123

tracker.com

Web Tracking

Cookie Syncing



GET advertiser.com/pixel.jpg

Response, Set-Cookie: User=userABC

advertiser.com

Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

tracker.com

advertiser.com

Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com

tracker.com

advertiser.com

Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com

GET syncID=user123, cookie=userABC

tracker.com

advertiser.com

Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com

GET syncID=user123, cookie=userABC

tracker.com

advertiser.com

- Third-parties with cookie syncing is enabled on 78% of modern websites :(

Defenses

- How might we defend against all these attacks?
- How can you defend against canvas fingerprinting?
- How can you defend against evercookies?
- How can you defend against cookie syncing?

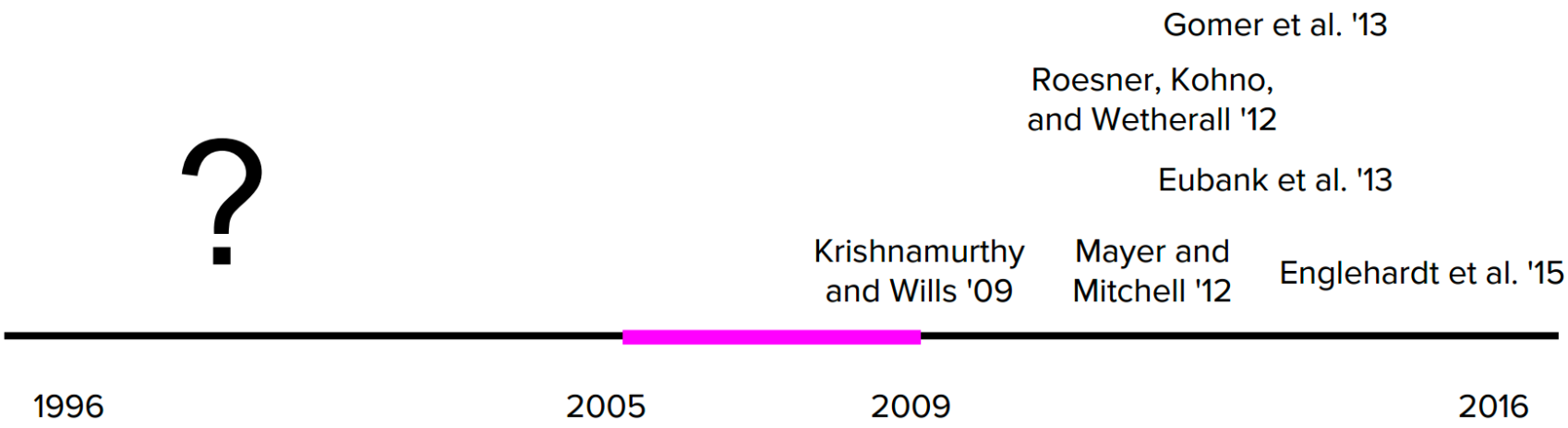
Some tools you should know about

Internet Jones and the Raiders of the Lost Trackers

This paper…

- What is the basic premise of the paper?
- How do the authors seek to evaluate their premise?
- What do the authors hope come out of their work?

The basic gist…



Wayback machine to the rescue!

- What is the wayback machine?
- What parts of the webpage does the wayback machine archive?

Wayback machine to the rescue!

- What is the wayback machine?
- What parts of the webpage does the wayback machine archive?
  - JavaScript, stylesheets, all resources that it can identify statically from the site contents (no dynamism)

Wayback machine to the rescue!

Deepak Kumar

Hi! I am a second year computer science PhD student at the University of Illinois, Urbana-Champaign. I work with Michael Bailey in the NSRG. I'm broadly interested in computer security and privacy, through the lens of Internet-wide measurements. Specifically, I'm interested in investigating how adversaries abuse Internet services to cause undue harm to end-users. This summer (2018), I'll be at Google interning on the Internet abuse research team.

I spend a good amount of time working on creative writing projects. You can learn more about that here.

You can reach me via email, Twitter, LinkedIn or in person on campus. Some thoughts are here. Other loosely organized thoughts are here. My CV.

Publications

Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, Michael Bailey. Skill Squatting Attacks on Amazon Alexa. To appear in 27th USENIX Security Symposium (USENIX Security '18), Baltimore, Maryland, August 15-17, 2018. [pdf]

Yi Zhou, Deepak Kumar, Surya Bakshi, Joshua Mason, Andrew Miller, Michael Bailey. Erays: Reverse Engineering Ethereum's Opaque Smart Contracts. To appear in 27th USENIX Security Symposium (USENIX Security '18), Baltimore, Maryland, August 15-17, 2018. [pdf]

Dave (Jing) Tian, Nolen Scaife, Deepak Kumar, Michael Bailey, Adam Bates, Kevin R. Butler. SoK: "Plug & Pray" Today – Understanding USB Insecurity in Versions 1 through C. In 39th IEEE Symposium on Security and Privacy (Oakland '18), San Fransico, CA. May 21-23, 2018. [pdf]

Deepak Kumar, Zhengping Wang, Matthew Hyder, Joseph Dickinson, Joshua Mason, Michael Bailey, Gabrielle Beck, David Adrian, Zakir Durumeric, J. Alex Halderman. Tracking Certificate Misissuance in the Wild. In 39th IEEE Symposium on Security and Privacy (Oakland '18), San Francisco, CA. May 21-23, 2018. [pdf]

TrackingExcavator

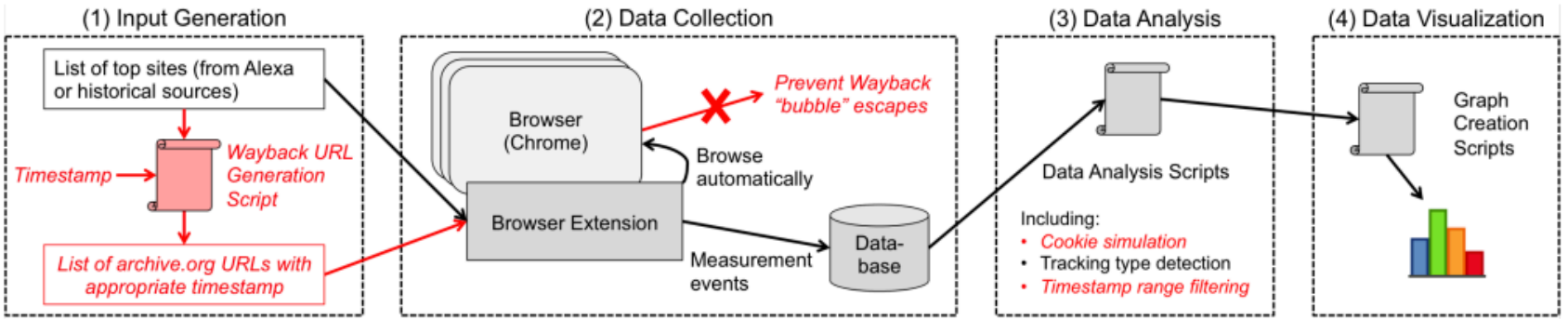


Figure 2: Overview of our infrastructure, TrackingExcavator, organized into four pipeline stages. Red/italic elements apply only to "Wayback mode" for historical measurements, while black/non-italics elements apply also to present-day measurements.

Turns out this is really hard…

- Some pages are not archived
- Some pages ask to not be archived
- Wayback escapes
  - WBM does not execute JS, but it might store JS that executes dynamically

Turns out this is really hard…

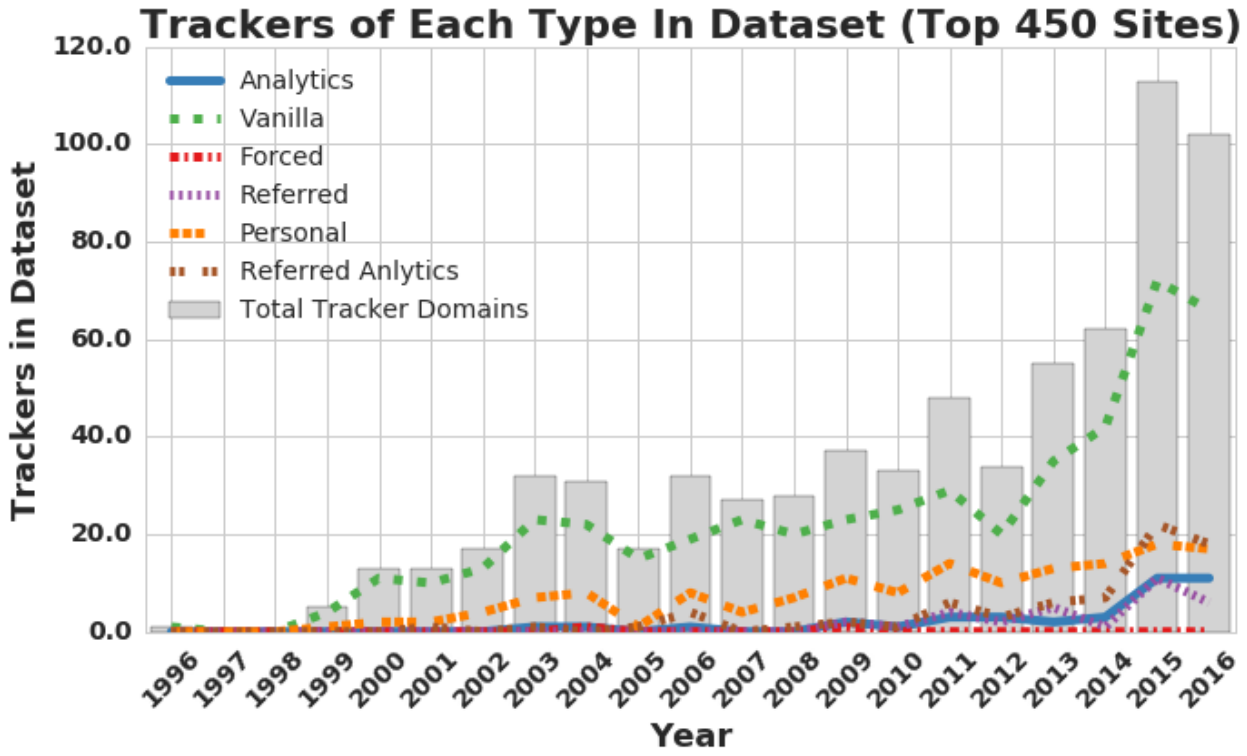
2008

2012

From https://web.archive.org/web/20151205072445/http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html

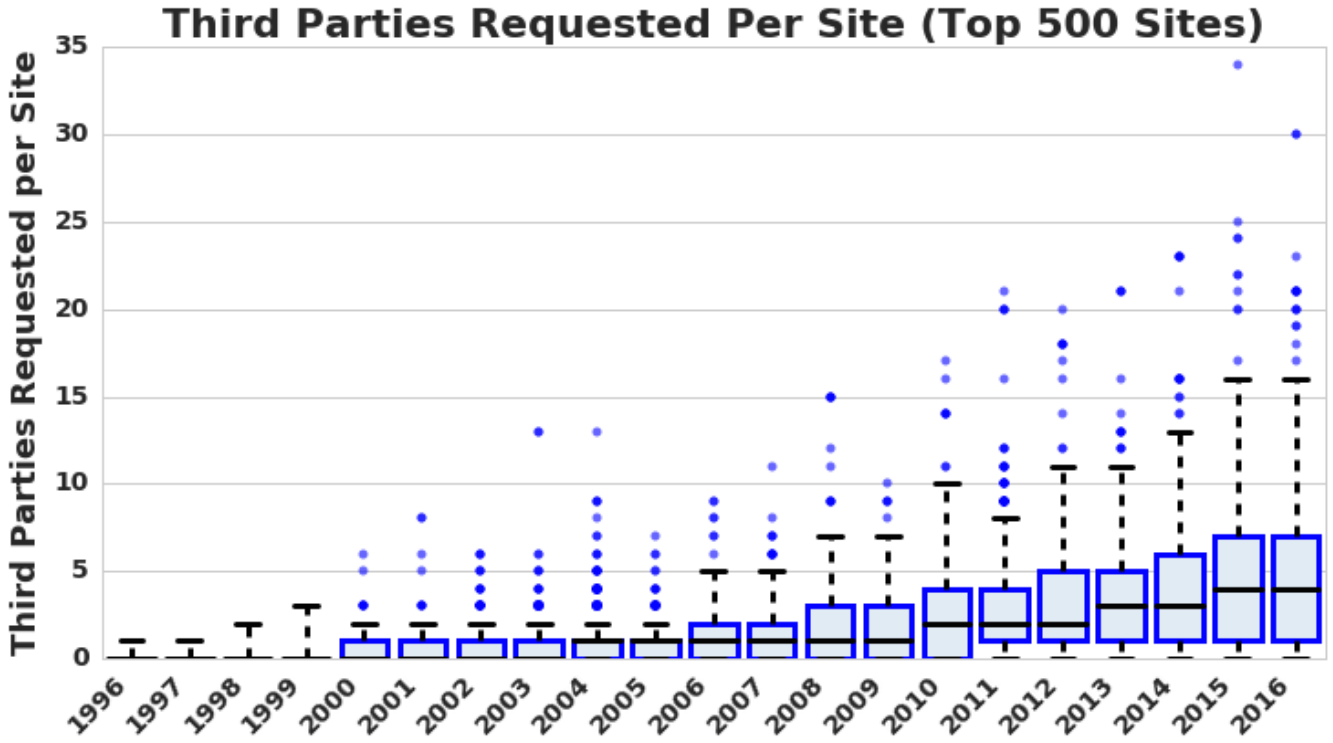
1996 – 2016: Tracking Analysis

More trackers of more types



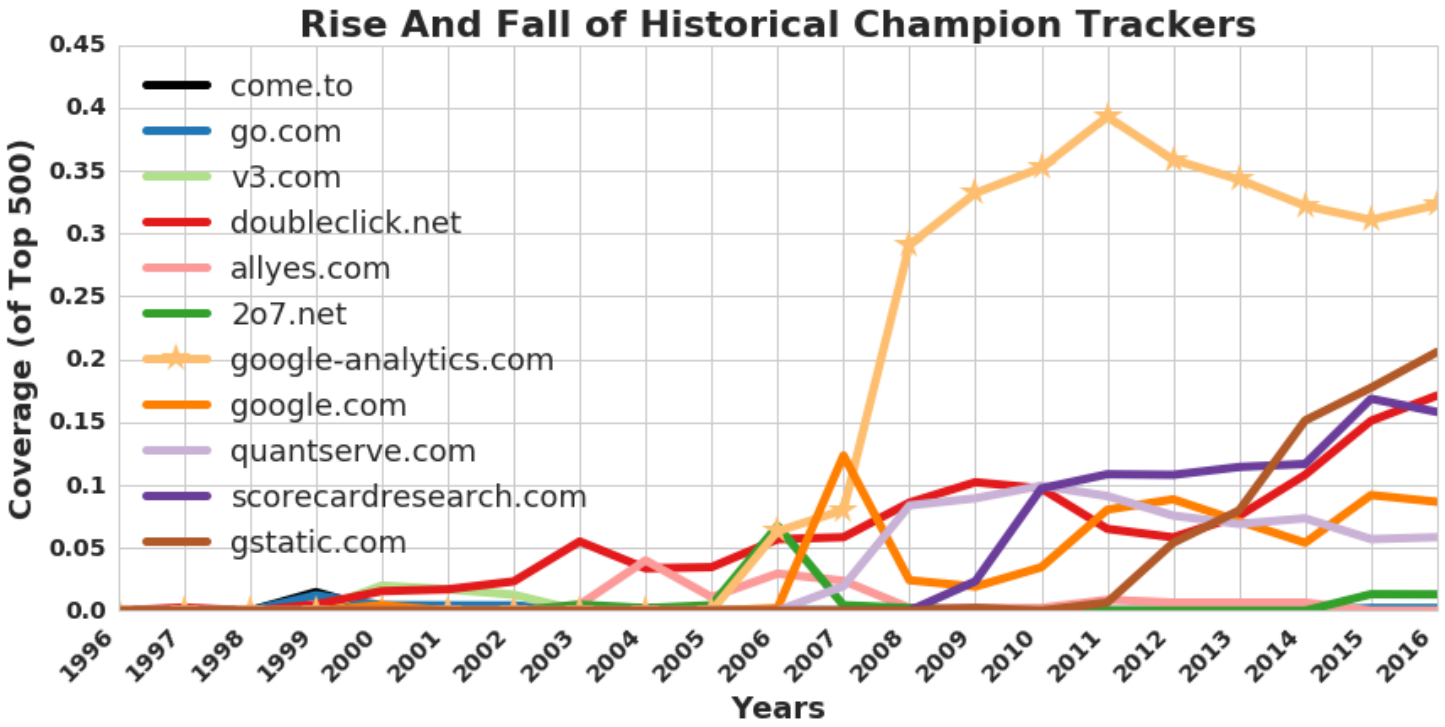
1996 – 2016: Tracking Analysis

More trackers of more types, **more per site**



1996 – 2016: Tracking Analysis

More trackers of more types, more per site, **more coverage**



Meta-thoughts on the paper

- What do we think about this paper? Did we enjoy it, why or why not?
- What are some limitations of the study?
- What are some practical things we can do with a study like this?

Next time…

- Moving away from the general “web” and talking about TLS – the mechanism that enables **HTTPS**
- I’ve made “The security of HTTPS Interception” **optional** reading