CSE227 – Graduate Computer Security

Web Fundamentals Catchup + Centralization I

UC San Diego

Housekeeping

General course things to know

- Course projects
  - Everyone should have received some feedback from me! **Please let me know if this was untrue…**
- Midterm check-in report is due **5/8 by EOD.** Don’t be late.
- I made the second paper for Thursday **optional**; so we can appropriately catch up!
- **Note:** No class next Tuesday; meet with your teams!

Project Management

Week 4 advice on projects

- For this week, if you're looking for something to "do…"
- Meet with your group, at least twice, and discuss the ideas you're finding. How do they help you think more clearly about your own project?
- Start to play with scaffolding code, repos, and resources. You should have a good idea, by the end of this week, **what exactly you need** in order to be successful
- **Ask me for resources by end of this week.** For example, if you need a small windows PC to test stuff, spec it out and make the request. Or if you need an API key for something, *ask in advance*.
    - You will not be successful if you don't know what you need to be successful

Today's lecture

Learning Objectives

- Recap some web concepts
- Discuss the HTML sanitization paper
- Introduce the basics of web tracking, and discuss the raiders paper; tee up web centralization

Web Recap

Remember from last time…

- What is a web client?
- What is a web server?
- When you make a request to a website, what *actually* happens?
- What's the same-origin policy, and what does it actually protect?
- What is a cross-site request forgery attack, and what's the Origin Header?

The Defense: Origin header

Origin

Baseline Widely available

The HTTP `Origin` request header indicates the origin (scheme, hostname, and port) that *caused* the request. For example, if a user agent needs to request resources included in a page, or fetched by scripts that it executes, then the origin of the page may be included in the request.

CSRF meta-questions

- How feasible is a CSRF attack? Will it work in practice?
- What software does the "Origin" proposal require you to *trust*?
  - Is this assumption always going to be true?
- How would you defend against a CSRF attack today? Is it that different from 2007, when this paper was written?
- What would you say is a **fundamental issue** that enables a CSRF attack?

CSRF meta-questions

- How feasible is a CSRF attack? Will it work in practice?
- What software does the “Origin” proposal require you to *trust*?
  - Is this assumption always going to be true?
- How would you defend against a CSRF attack today? Is it that different from 2007, when this paper was written?
- What would you say is a **fundamental issue** that enables a CSRF attack?
  - Side-effects in the interface between the web server and web browser
  - *Feature*, not a bug

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

What is HTML?

What is HTML?

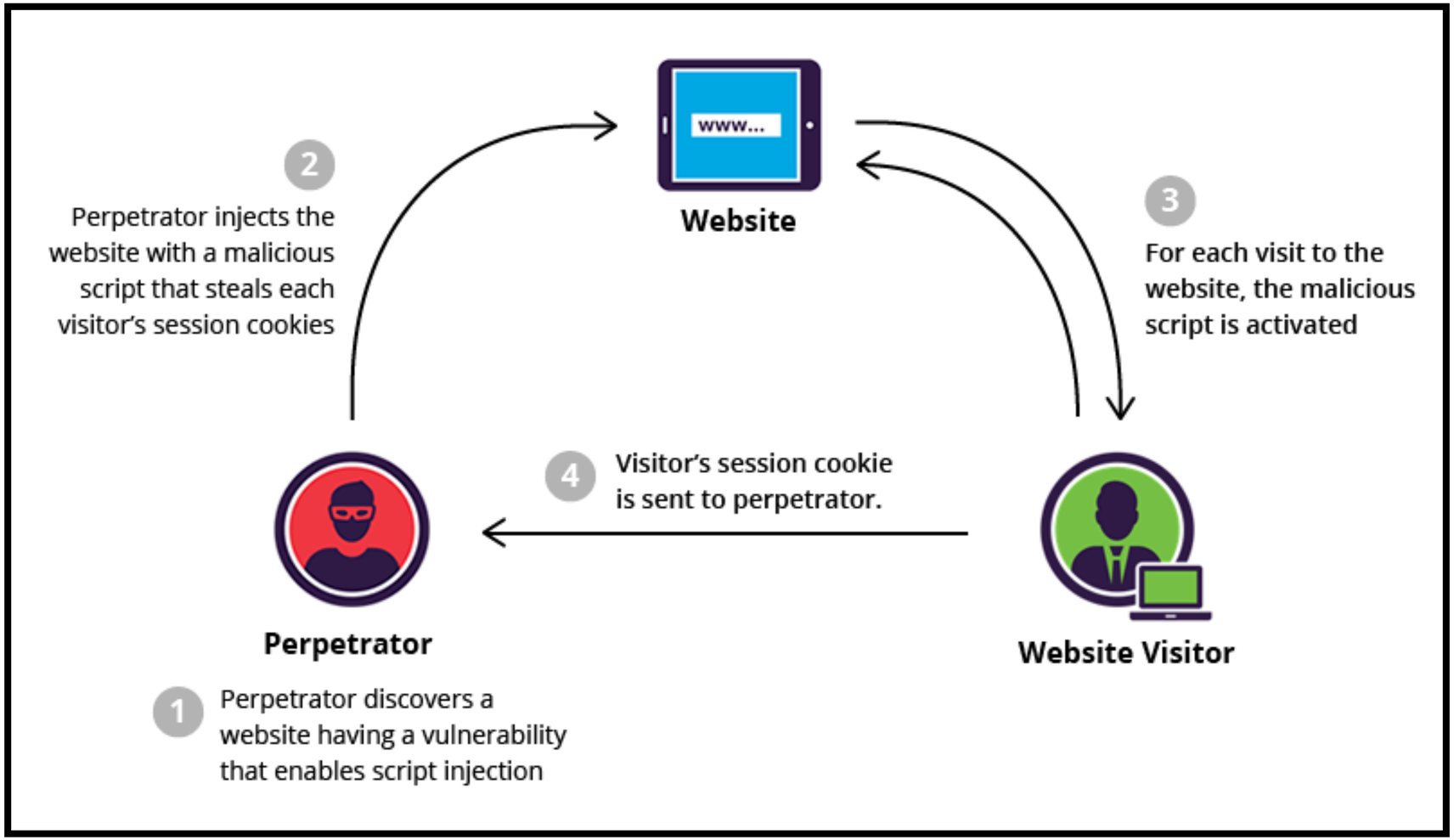
Hypertext Markup Language: The **structure** of how we embed web content into web pages.

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

How does XSS work?



Imperva

Very simple XSS example

Web Server

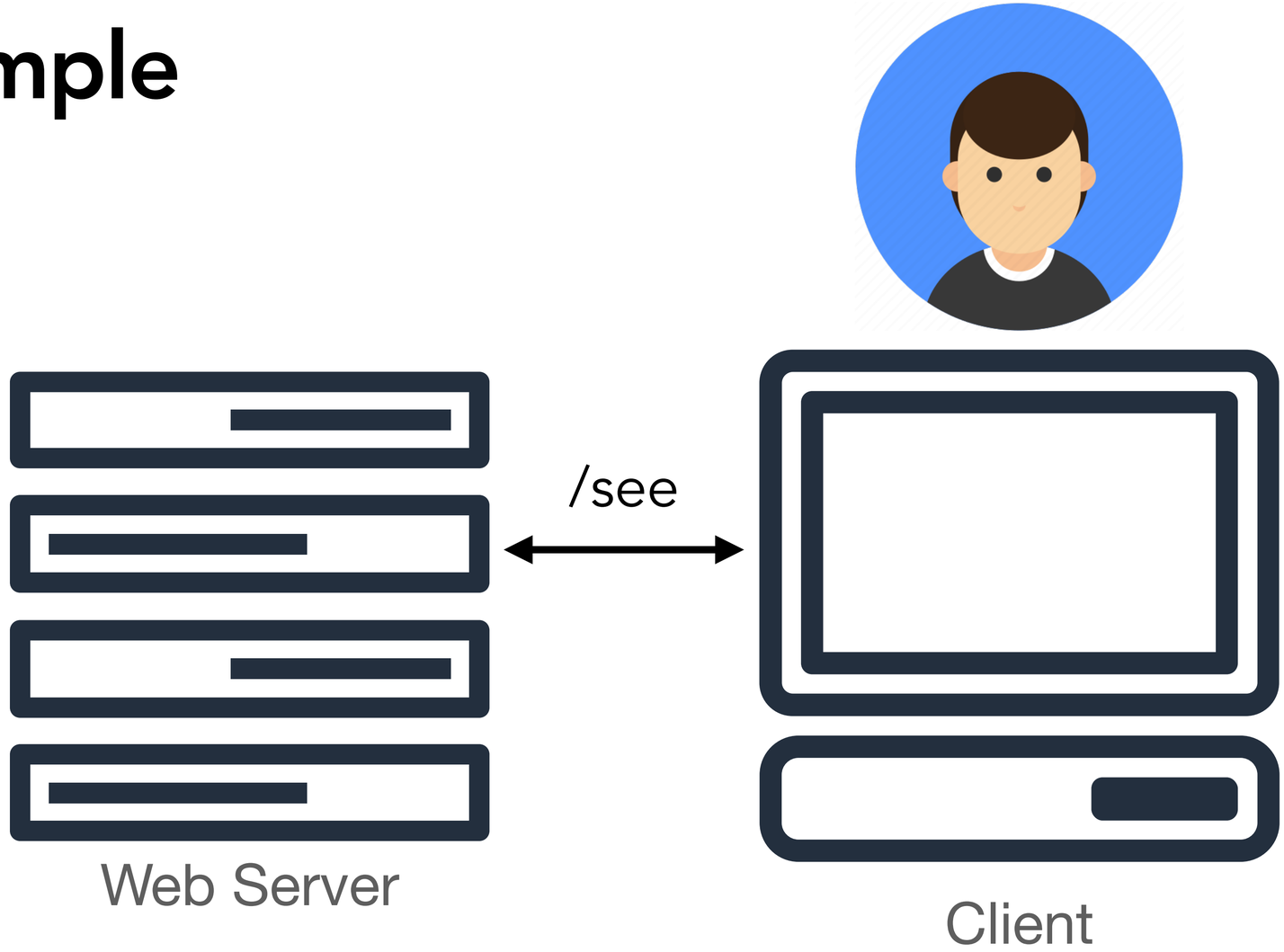
Client

Very simple XSS example

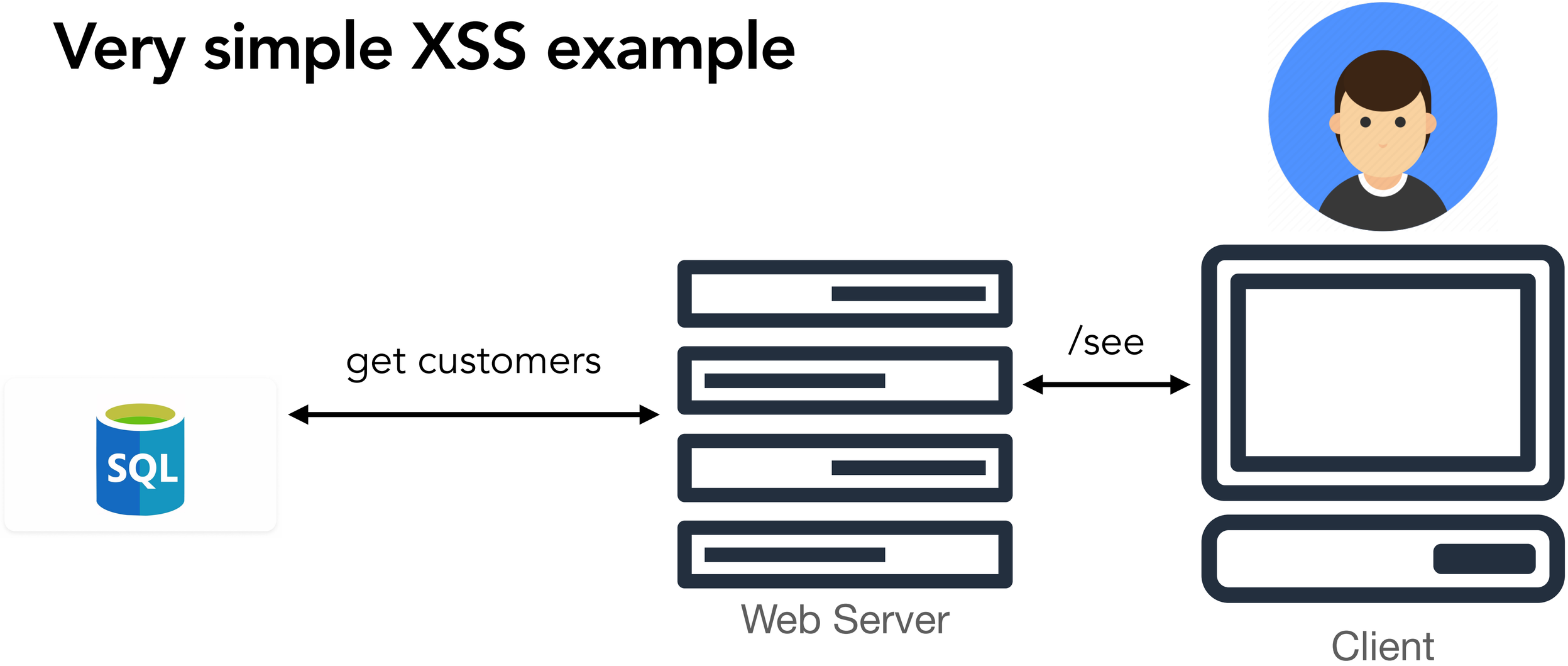
Web Server

Client

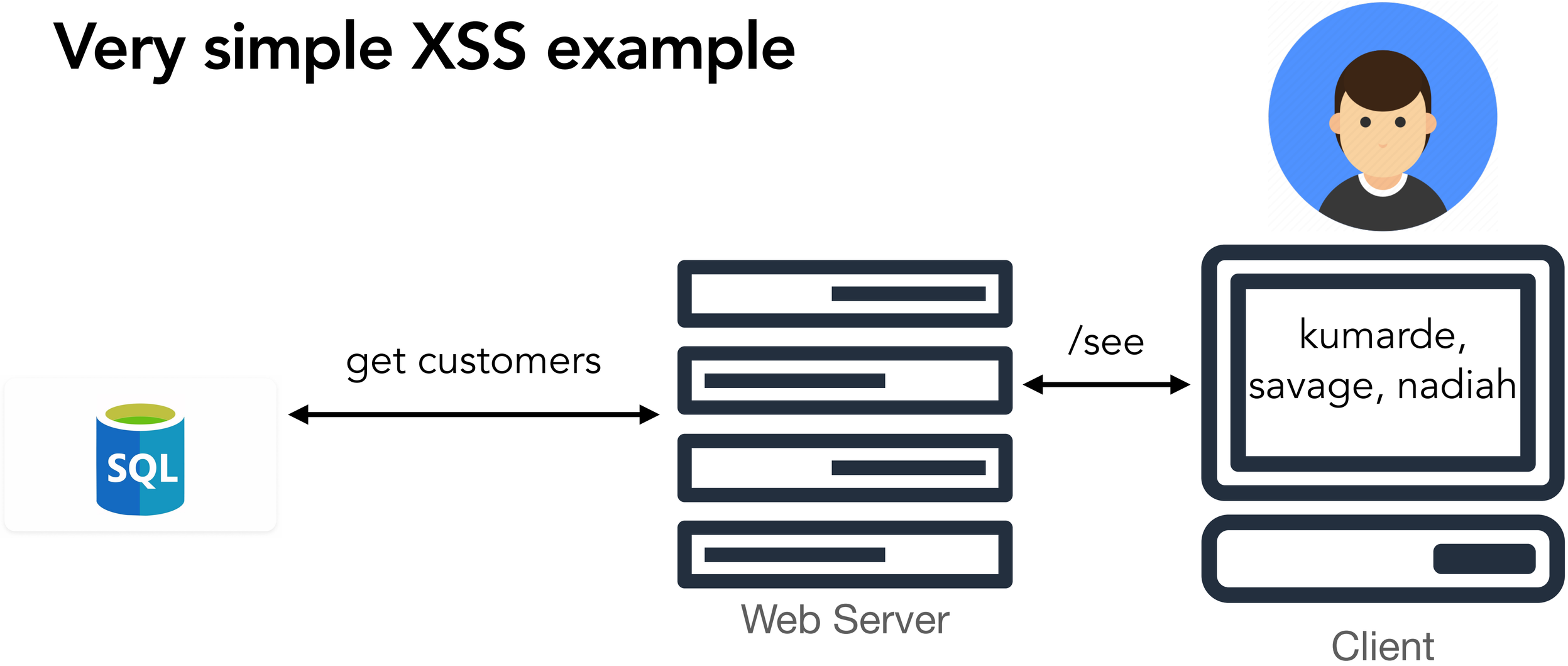
Very simple XSS example



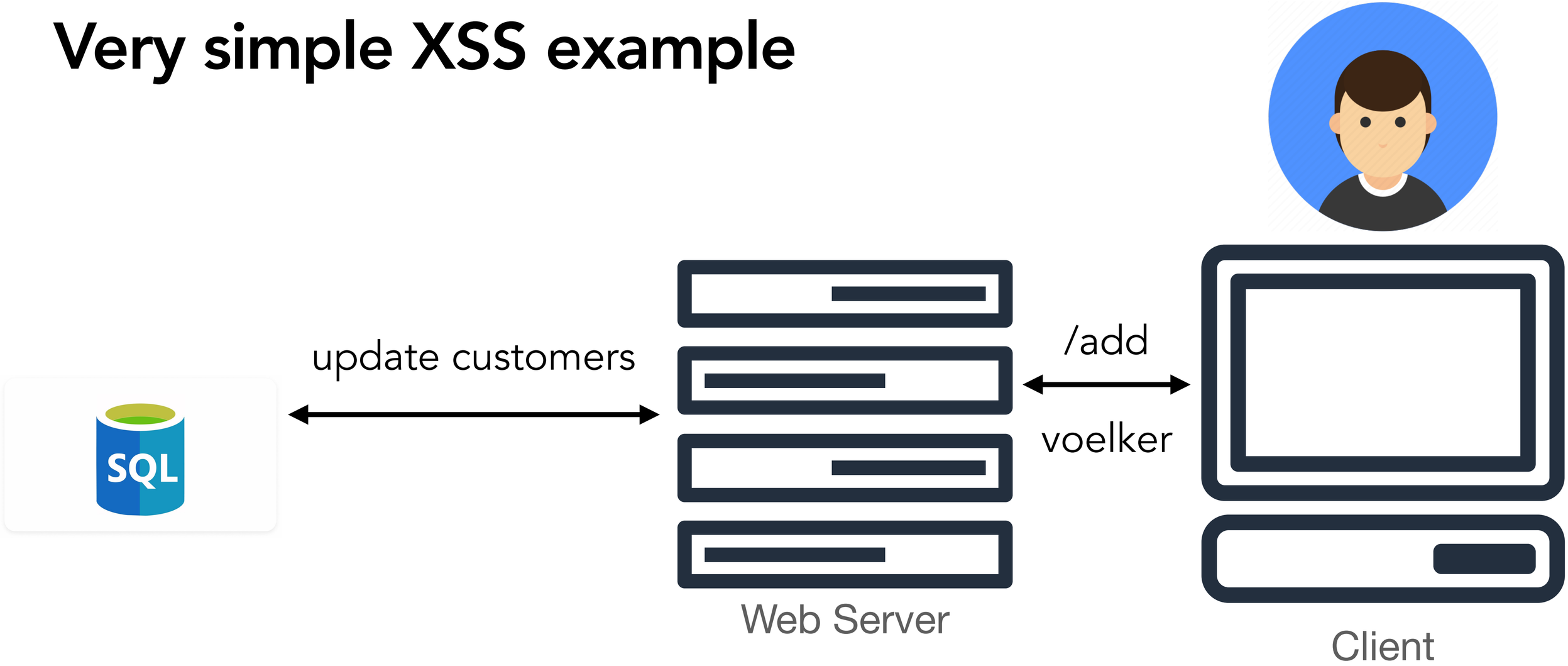
Very simple XSS example



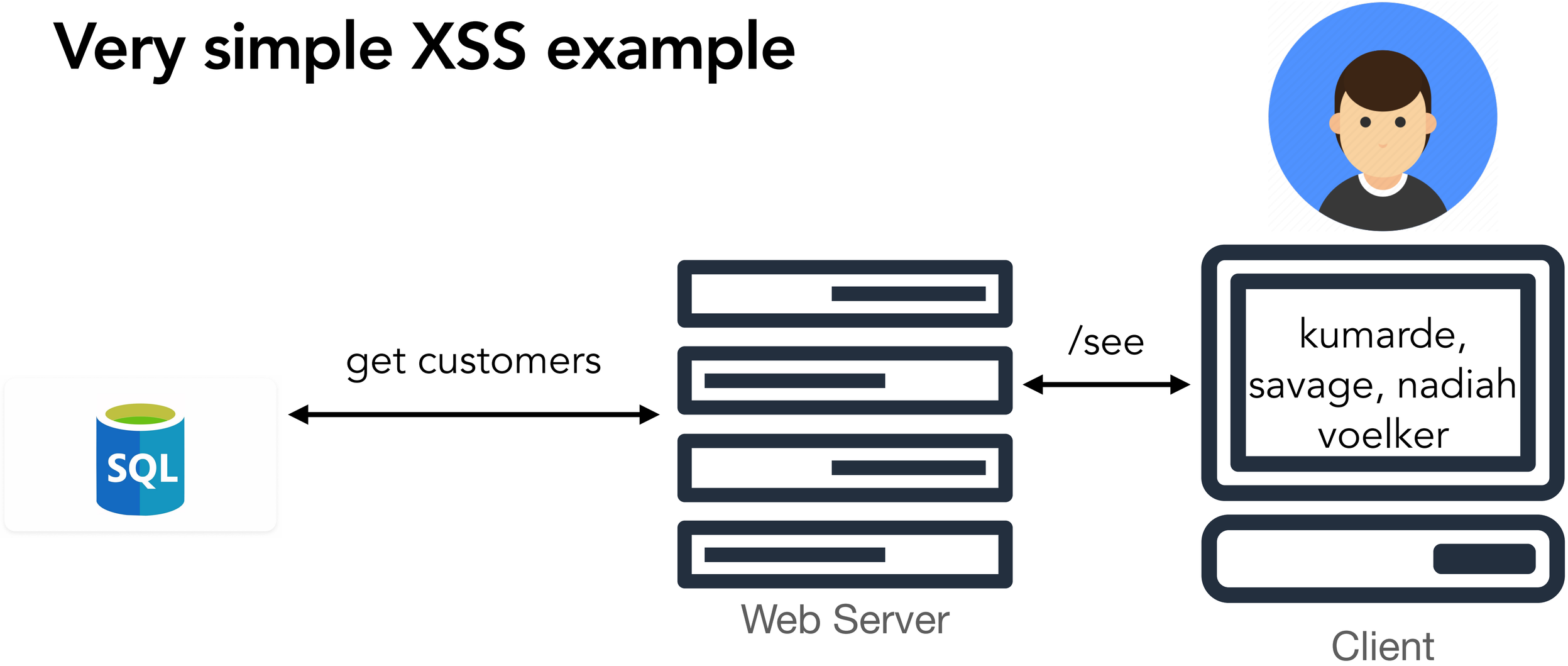
Very simple XSS example



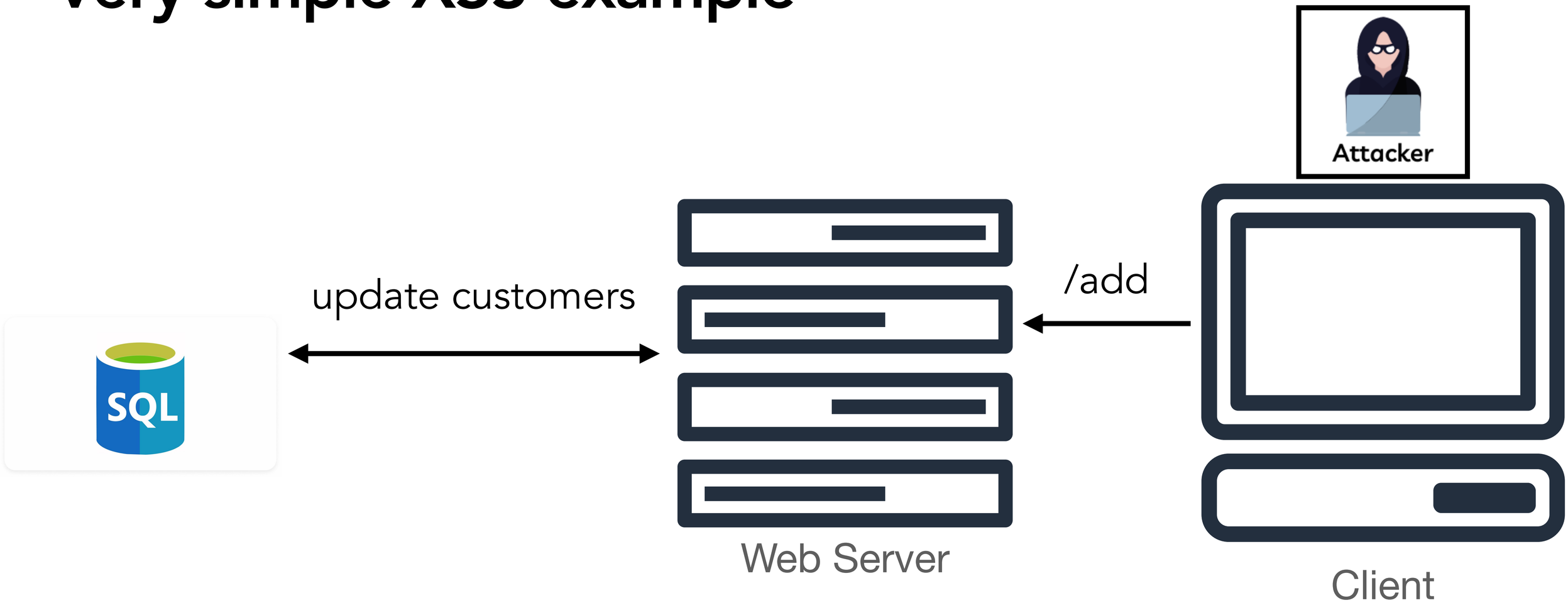
Very simple XSS example



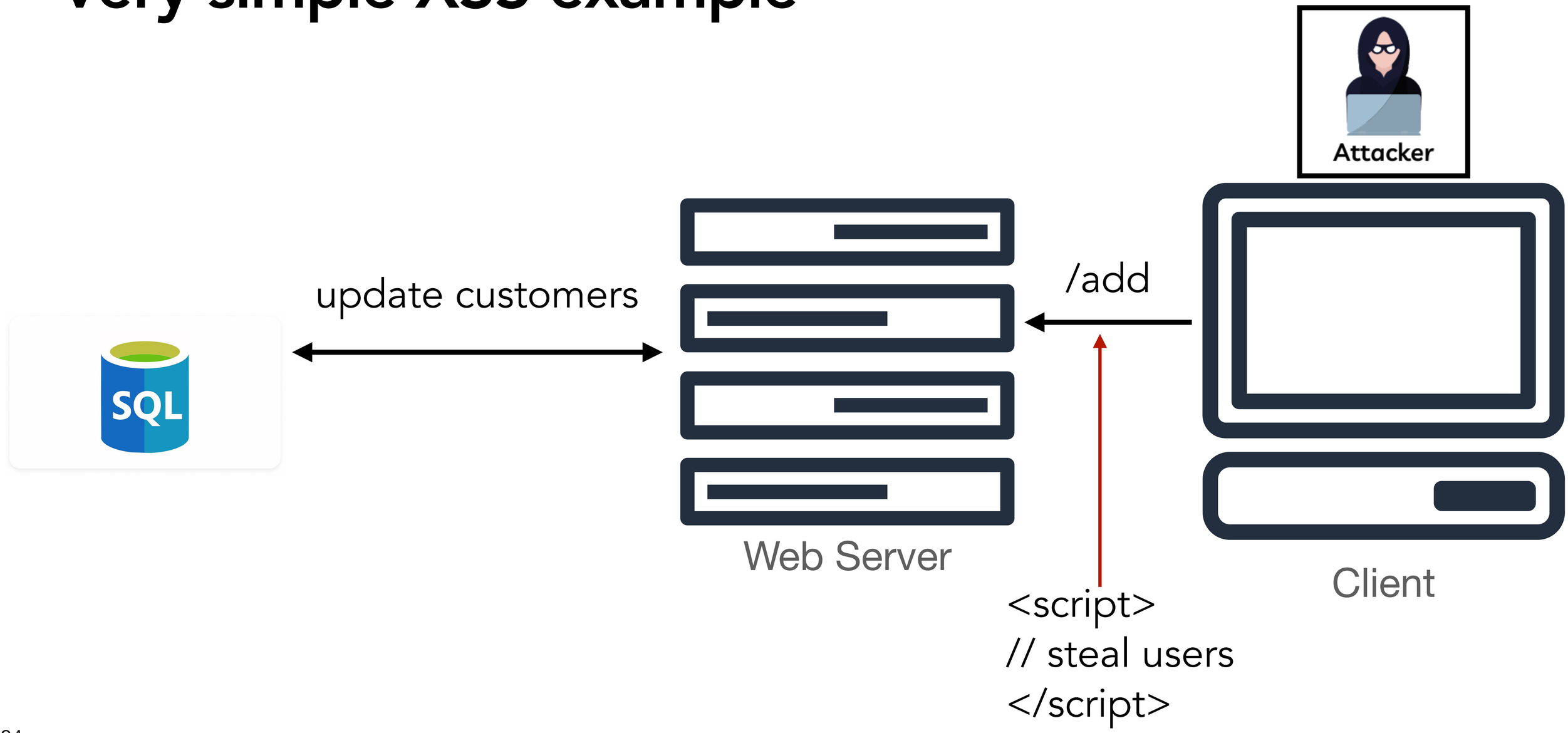
Very simple XSS example



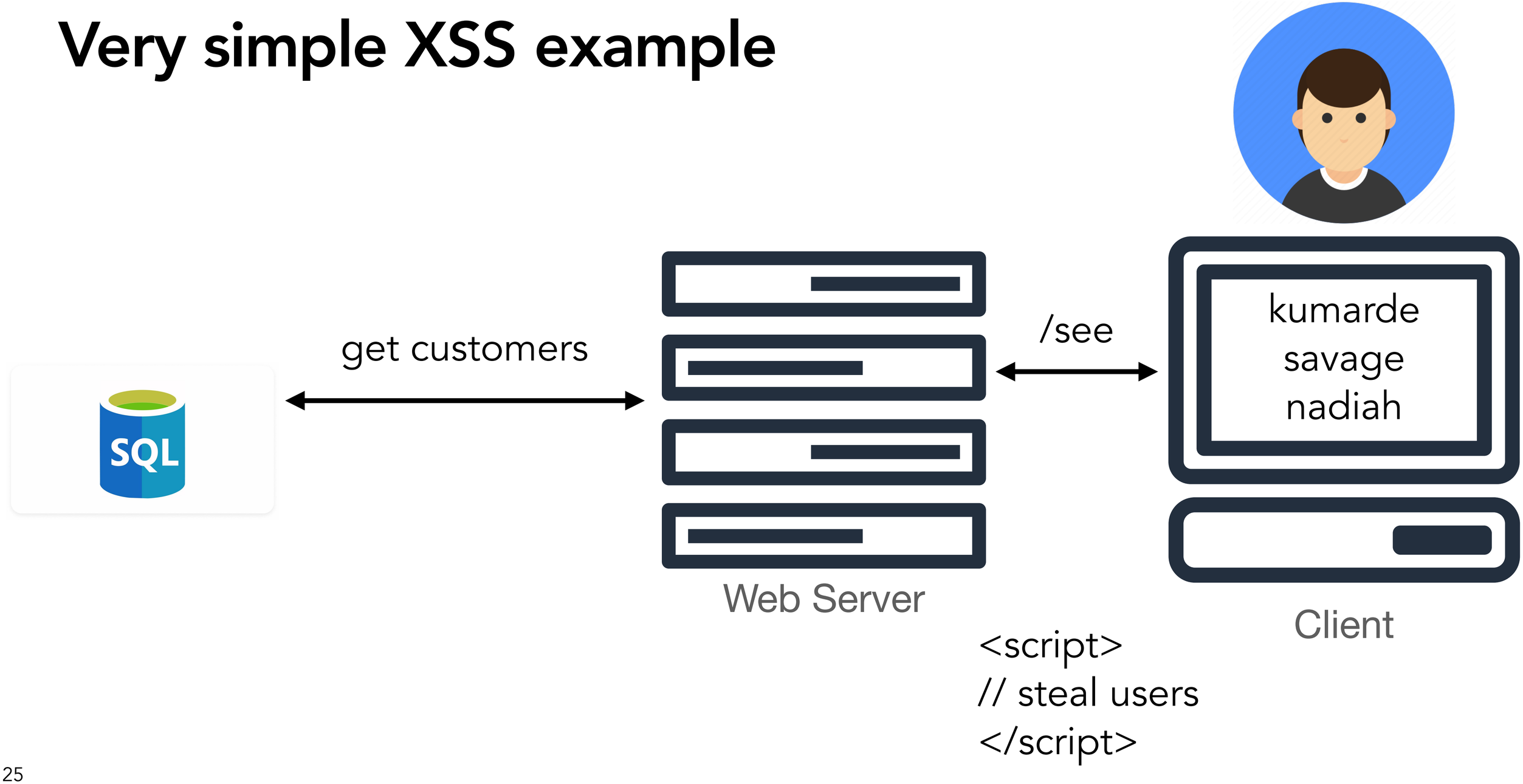
Very simple XSS example



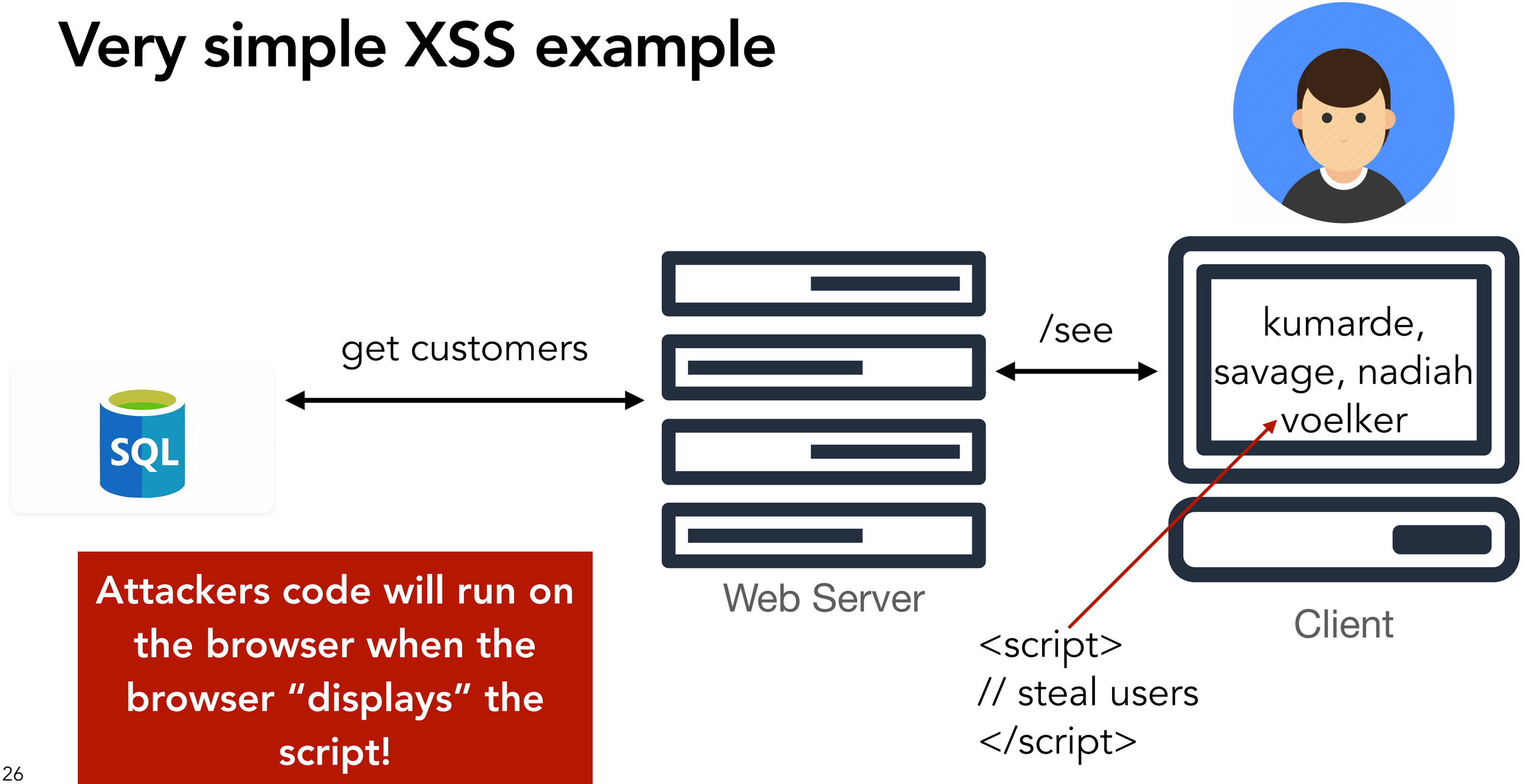
Very simple XSS example



Very simple XSS example



Very simple XSS example



Attackers code will run on the browser when the browser "displays" the script!

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

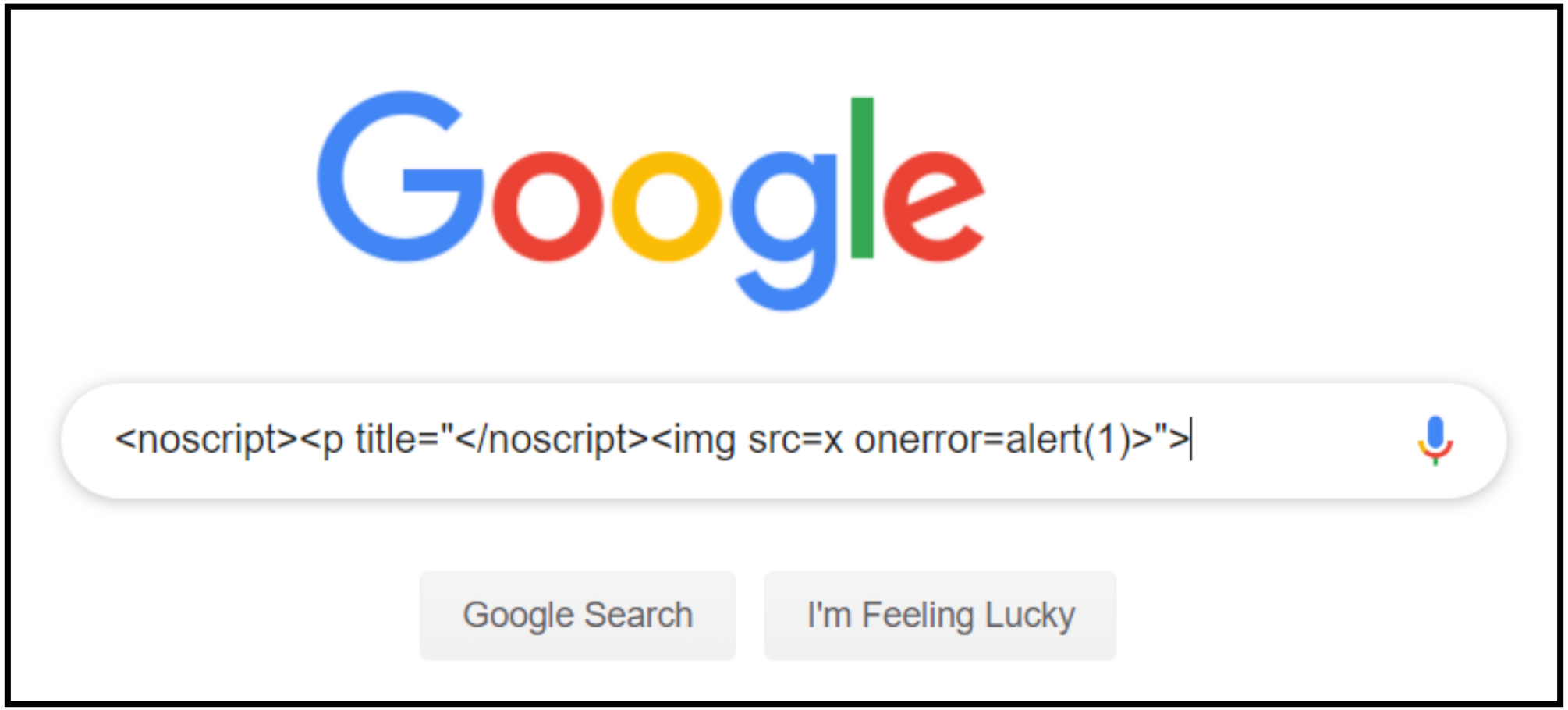
- Why is accurate HTML sanitization quite hard for servers to do?
  - Context dependent
  - Requires understanding how the browser is going to interpret the HTML, which turns out is not easy!

What is Mutation Cross-Site Scripting (mXSS)?

What is Mutation Cross-Site Scripting (mXSS)?

- Sanitizing some HTML may result in *new* HTML that in and of itself is vulnerable due to browser parsing implementation
  - In short, parsing / sanitization is… really hard

mXSS example: Google Search in 2019



mXSS example: Google Search in 2019

- Google used to rely on a client-side sanitization JavaScript library (they created) called **closure**
  - —> a client side sanitization library that Google used called DOMPurify used **closure**
- In 2018, a developer removed a small input sanitization check in **closure,** likely due to some conflict somewhere else in the library
- <noscript> is… painful to implement
  - Represents **nothing** if JS is not enabled, but represents its **children** if JS is enabled…

mXSS example: Google Search in 2019

- <noscript><p title="</noscript><img src="x" onerror=alert(1);>">

mXSS example: Google Search in 2019

- <noscript><p title="</noscript><img src="x" onerror=alert(1);>">
- This gets interpreted as…
  - <noscript><p title="</noscript>
    <img src="x" onerror="alert(1)">
    ''">"
- **The alert runs in the image tag! O snap!**
- Google was vulnerable for **5 months!** No one knows if this was used for malicious purposes, but our guess is probably not…

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

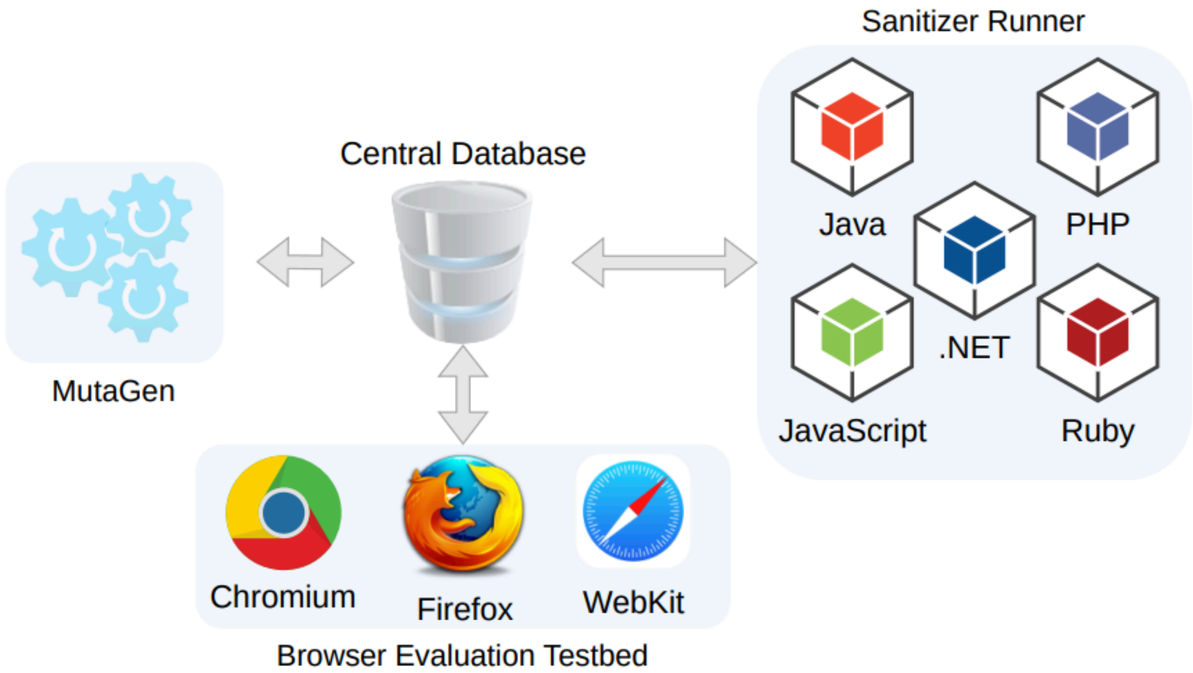


Figure 3: Sanitizer Evaluation Setup

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
- This paper was 2024…. IMO this type of mutation is a pretty good use case for LLMs…

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

Announcing their findings

- What is vulnerability disclosure?

Announcing their findings

- What is vulnerability disclosure?
- Who did the authors contact about the vulnerabilities they found?

Announcing their findings

- What is vulnerability disclosure?
- Who did the authors contact about the vulnerabilities they found?
  - Libraries, **and** open-source projects that use the vulnerable combination of DOMPurify + JSDom… e.g., Mozilla, Grafana, etc.
- What's the Google Caja story?

Announcing their findings

- What is vulnerability disclosure?
- Who did the authors contact about the vulnerabilities they found?
  - Libraries, **and** open-source projects that use the vulnerable combination of DOMPurify + JSDom… e.g., Mozilla, Grafana, etc.
- What's the Google Caja story?
- **Discussion question:** When is the right time to disclose a vulnerability you find?

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 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?
  - 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

Internet Jones and the Raiders of the Lost Trackers

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

Web Tracking Fundamentals

- What is *third-party* web tracking (as opposed to first-party web tracking)?
- How does third-party web tracking happen? Who allows these third-parties onto websites?
- What is cookie-based web tracking?
- What is fingerprint-based web tracking?

PODCAST: Tug of War | TRENDING: World Series | Pope and Biden | Tuesday elections | Halloween train attack | Economic bills | G20 summit | Box office |

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

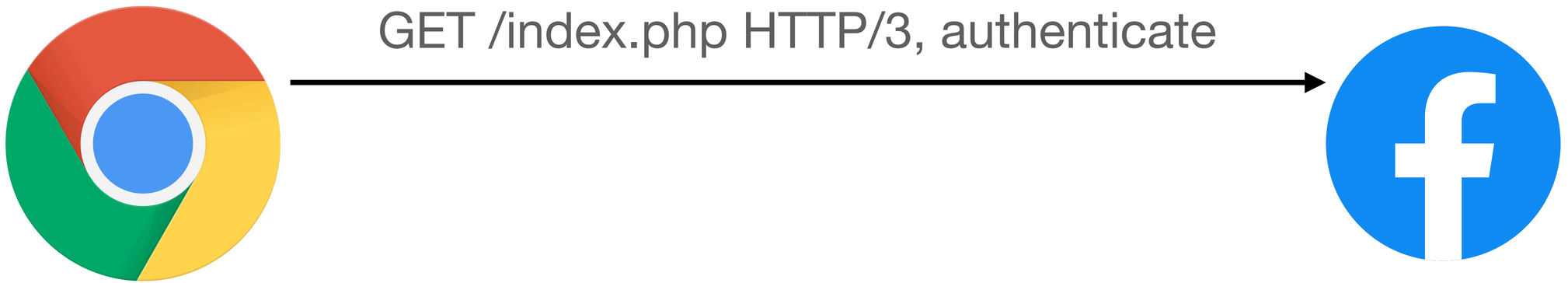
Cookies and Code

- 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

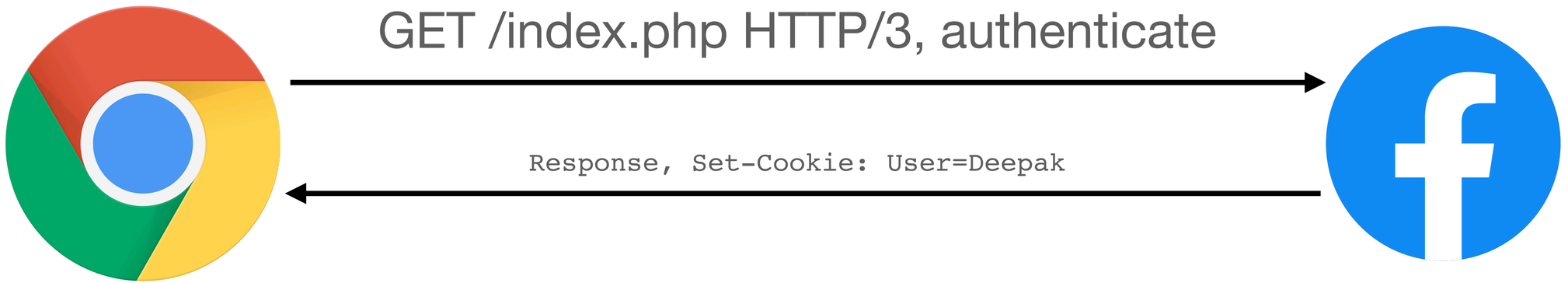
- Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)



Web Tracking

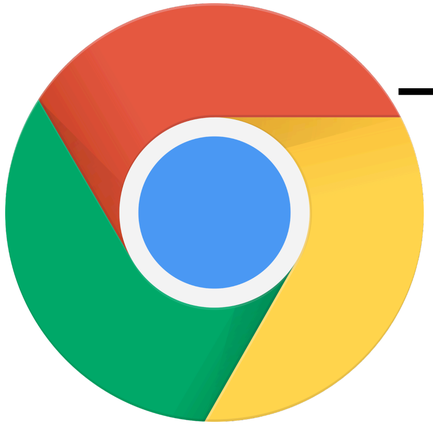
Cookies and Code

- 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.)

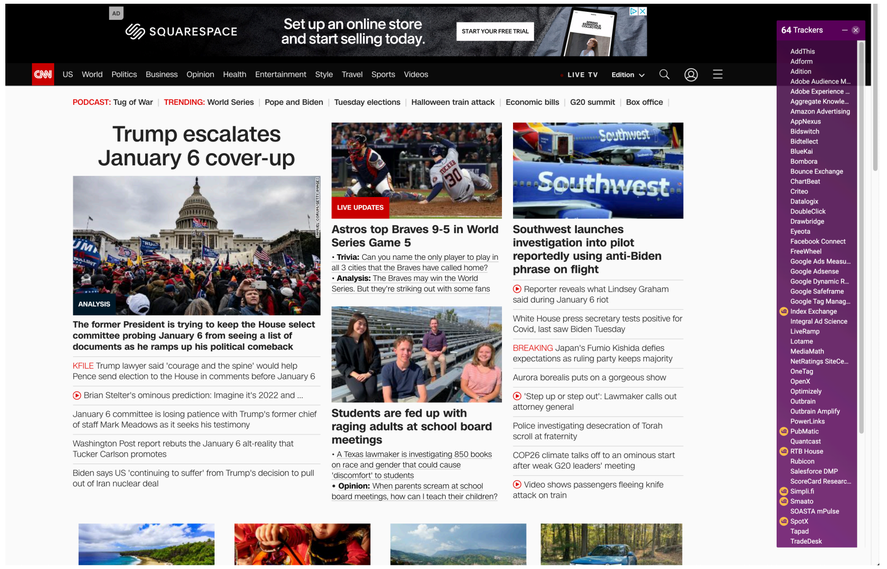


Web Tracking

Cookies and Code

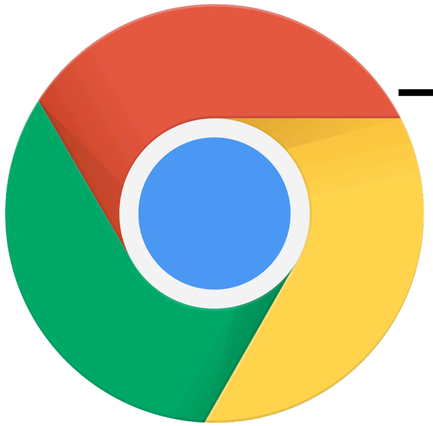


GET / HTTP/3



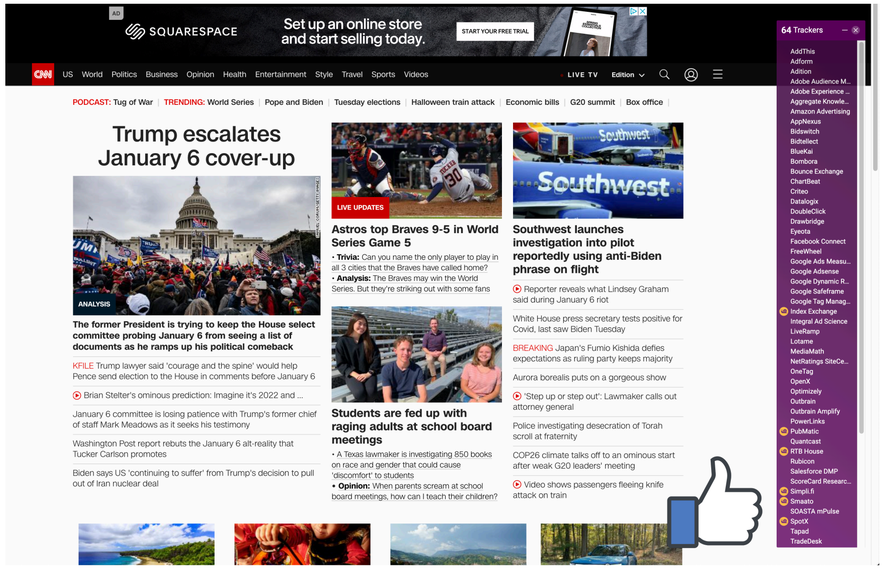
Web Tracking

Cookies and Code



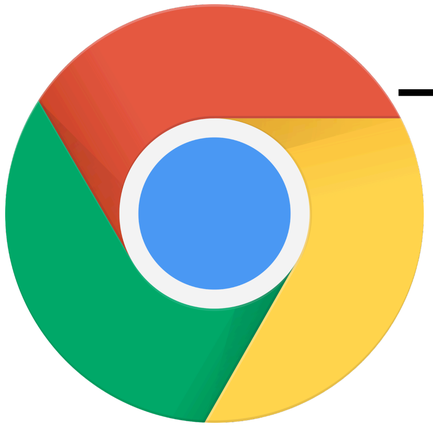
GET / HTTP/3

GET /facebook-like.js HTTP/3



Web Tracking

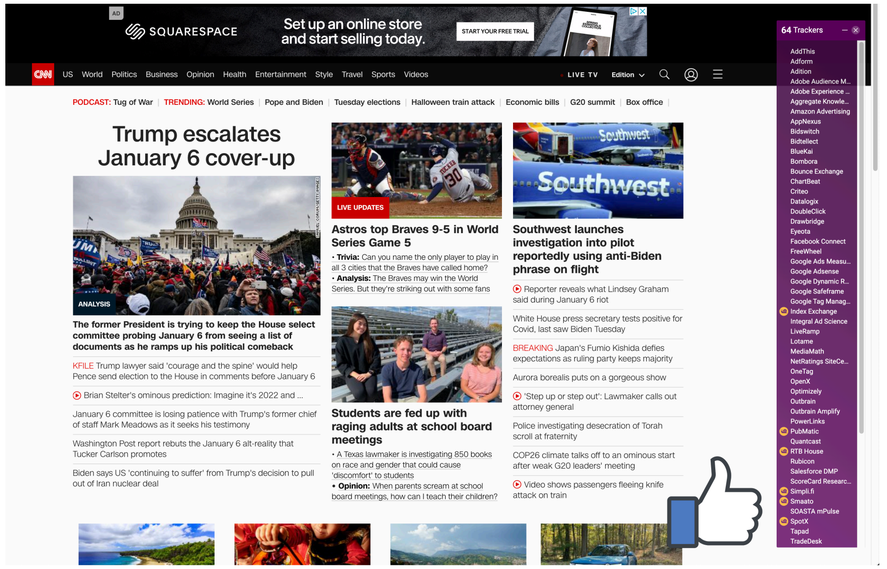
Cookies and Code



GET / HTTP/3

GET /facebook-like.js HTTP/3

Cookie: User=Deepak, Referer=cnn.com

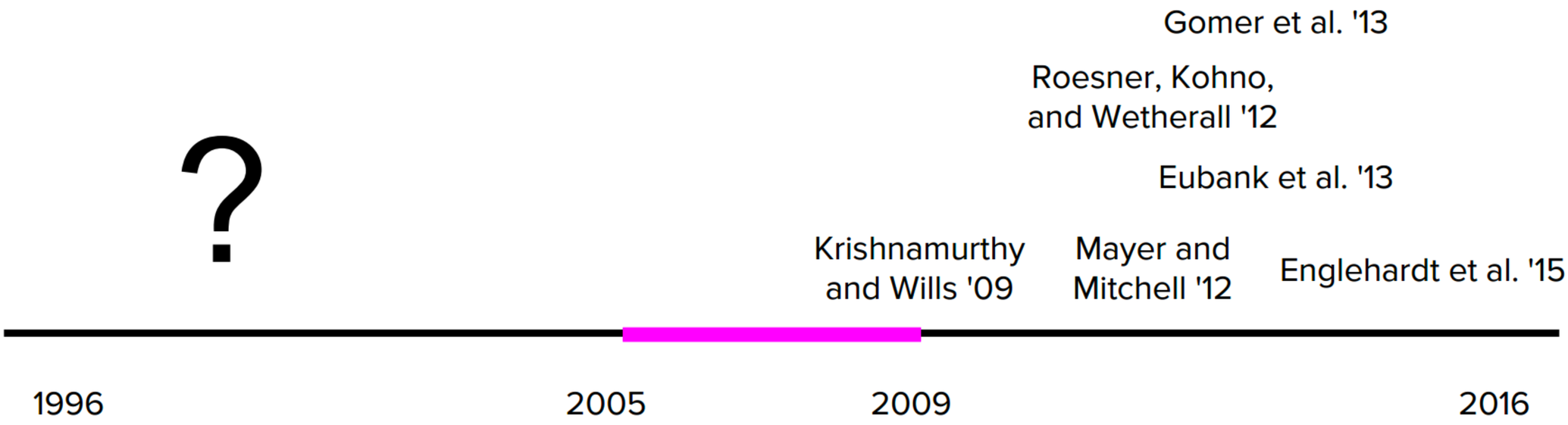


- With this request, companies can link your cookie to your browsing data (e.g., through Referer header, Host headers, Origin, or just JavaScript)

This paper’s goals

- 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

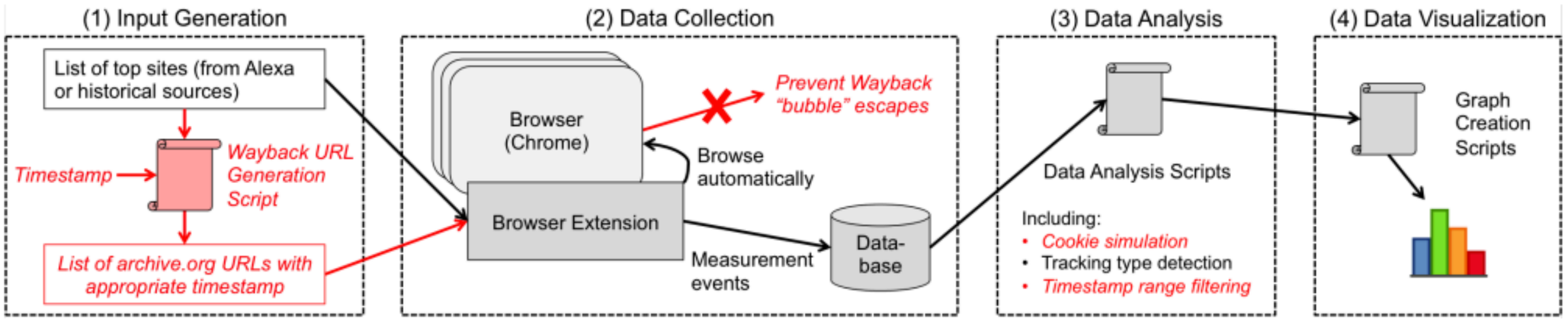


Figure 2: Overview of our infrastructure, TrackingExcavator, organized into four pipeline stages. Red/italic elements apply only to "Wayback mode" for historical measurements, while black/non-italics elements apply also to present-day measurements.

Defining things

- How do the authors *identify* tracking behaviors?

Defining things

- How do the authors *identify* tracking behaviors?
  - Use an existing taxonomy [60] and implement checks for the types of behaviors outlined… **is this foolproof?**

Defining things

- How do the authors *identify* tracking behaviors?
  - Use an existing taxonomy [60] and implement checks for the types of behaviors outlined… **is this foolproof?**
- What websites do the authors use as their *measurement set*?
  - What's the assumption baked into this choice of domains?

Turns out, Wayback Machine API 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
- Archival cookies need to be logged and replayed, which requires significant effort…
  - Still a challenge today!

Turns out this is really hard…

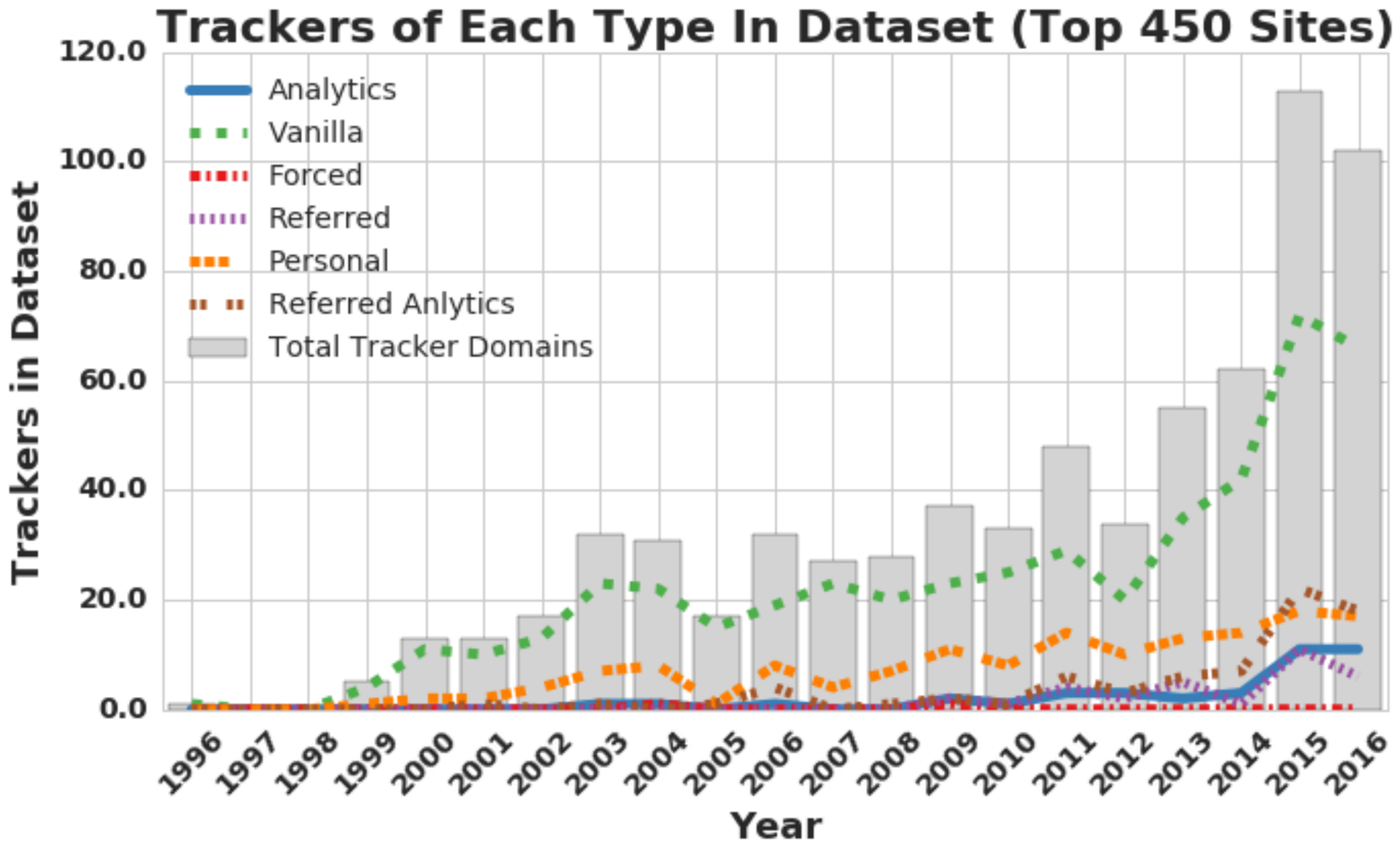
2008

2012

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

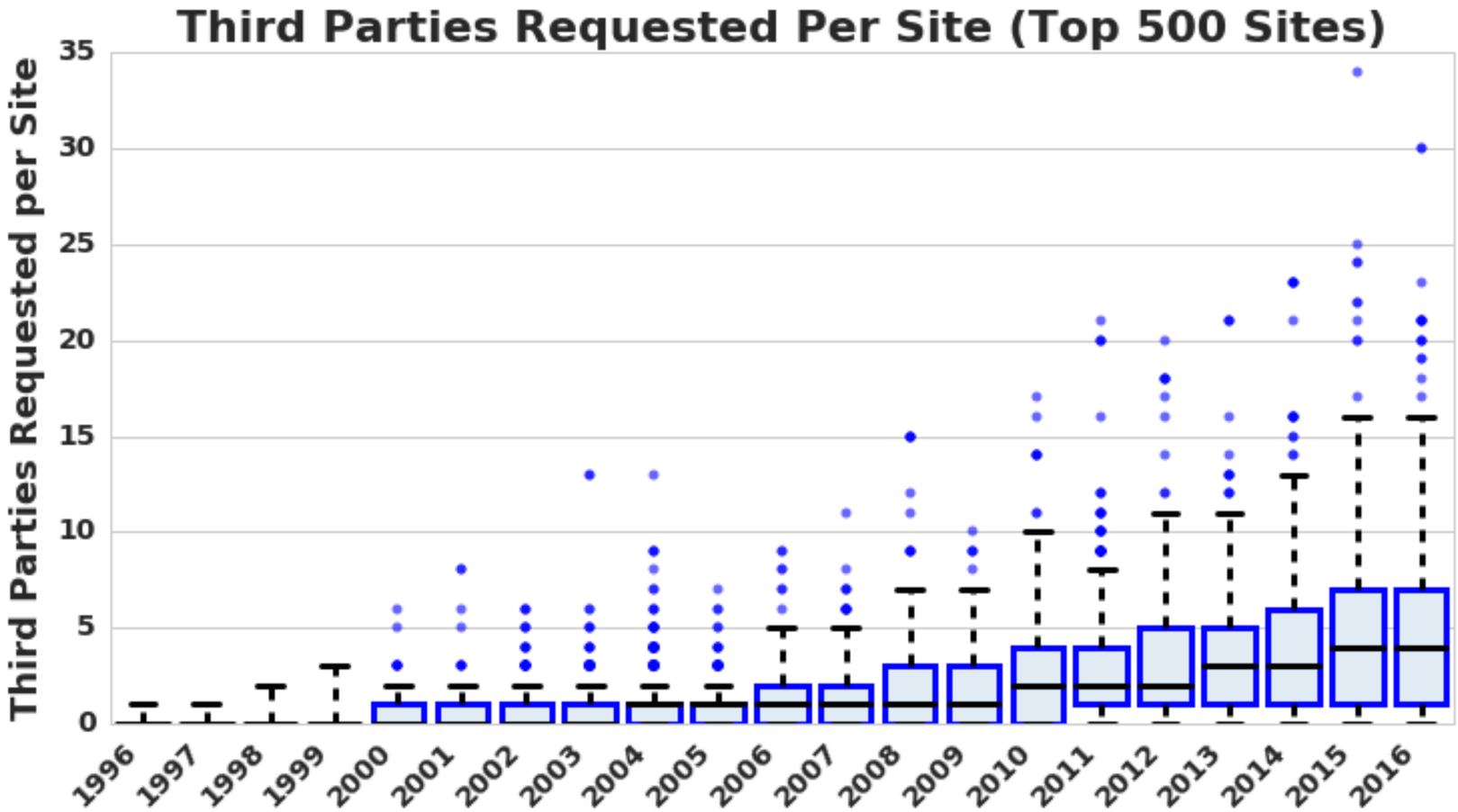
1996 – 2016: Tracking Analysis

More trackers of more types



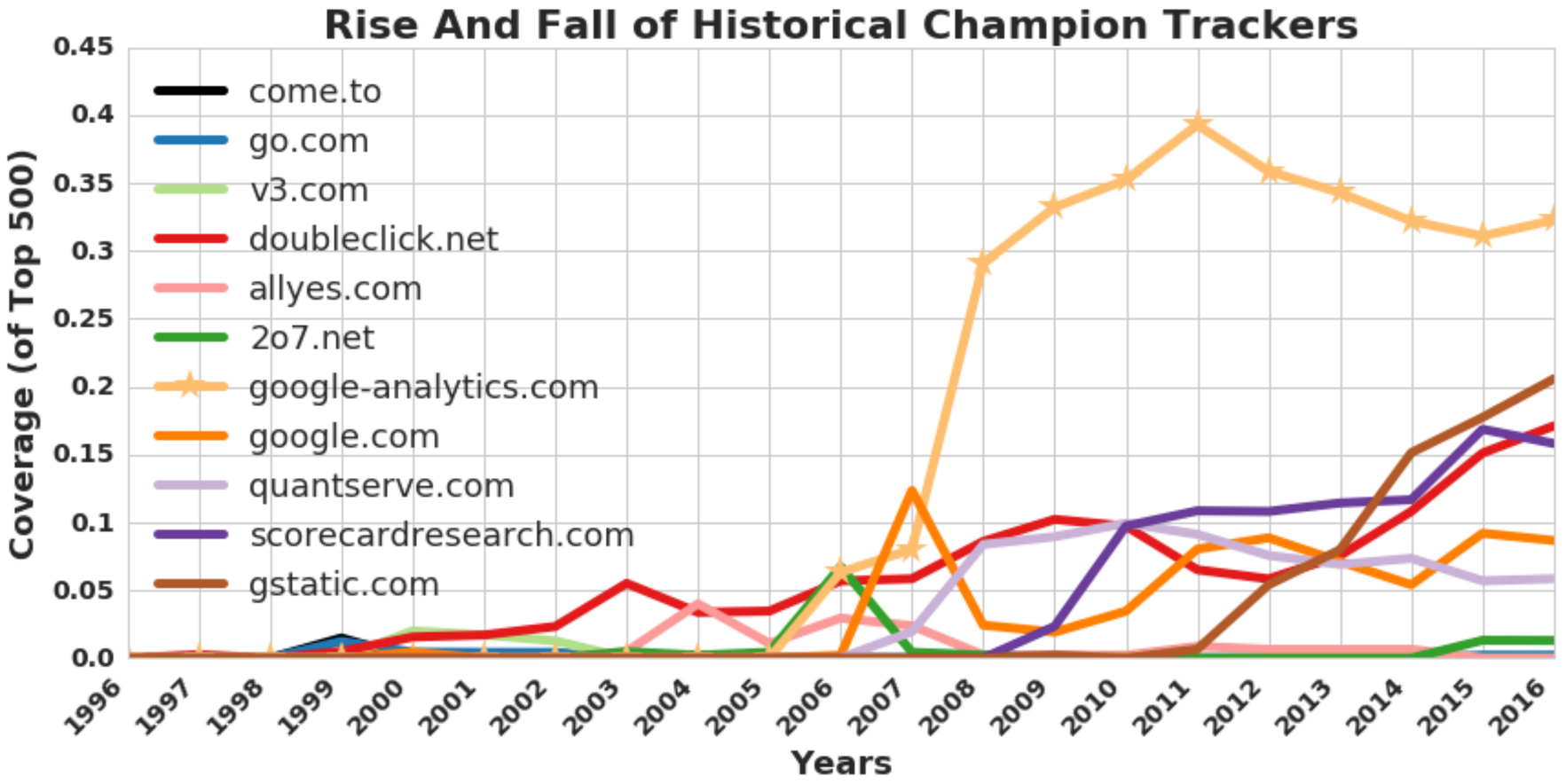
1996 – 2016: Tracking Analysis

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



1996 – 2016: Tracking Analysis

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



The most prominent trackers

- Who is the most prominent tracker on the Internet?

The most prominent trackers

- Who is the most prominent tracker on the Internet?
  - Google! Through Google Analytics.
- Today, **>70% of Top 1M websites load Google analytics**
  - Increased reliance on a small handful of players: Google, Facebook, Amazon…
- Today's language: hyperscalar. What's a hyperscalar?

Domain	Top 1M	Domain	Top 1M
google-analytics.com	67.8%	ajax.googleapis.com	23.1%
gstatic.com	50.1%	googlesyndication.com	19.6%
fonts.googleapis.com	42.8%	googleadservices.com	14.1%
doubleclick.net	40.5%	twitter.com	12.8%
facebook.com	33.7%	fbcdn.net	10.7%
google.com	33.2%	adnxs.com	10.5%
facebook.net	27.4%		

Discussion: On Centralization…

- Is centralization good for the Internet or bad for the Internet?
- What kinds of *benefits* would you want from centralization? What kinds of *downsides* can you see from increased centralization?

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…

- Continuing with web centralization theme… and also introducing **TLS** and **HTTPS**