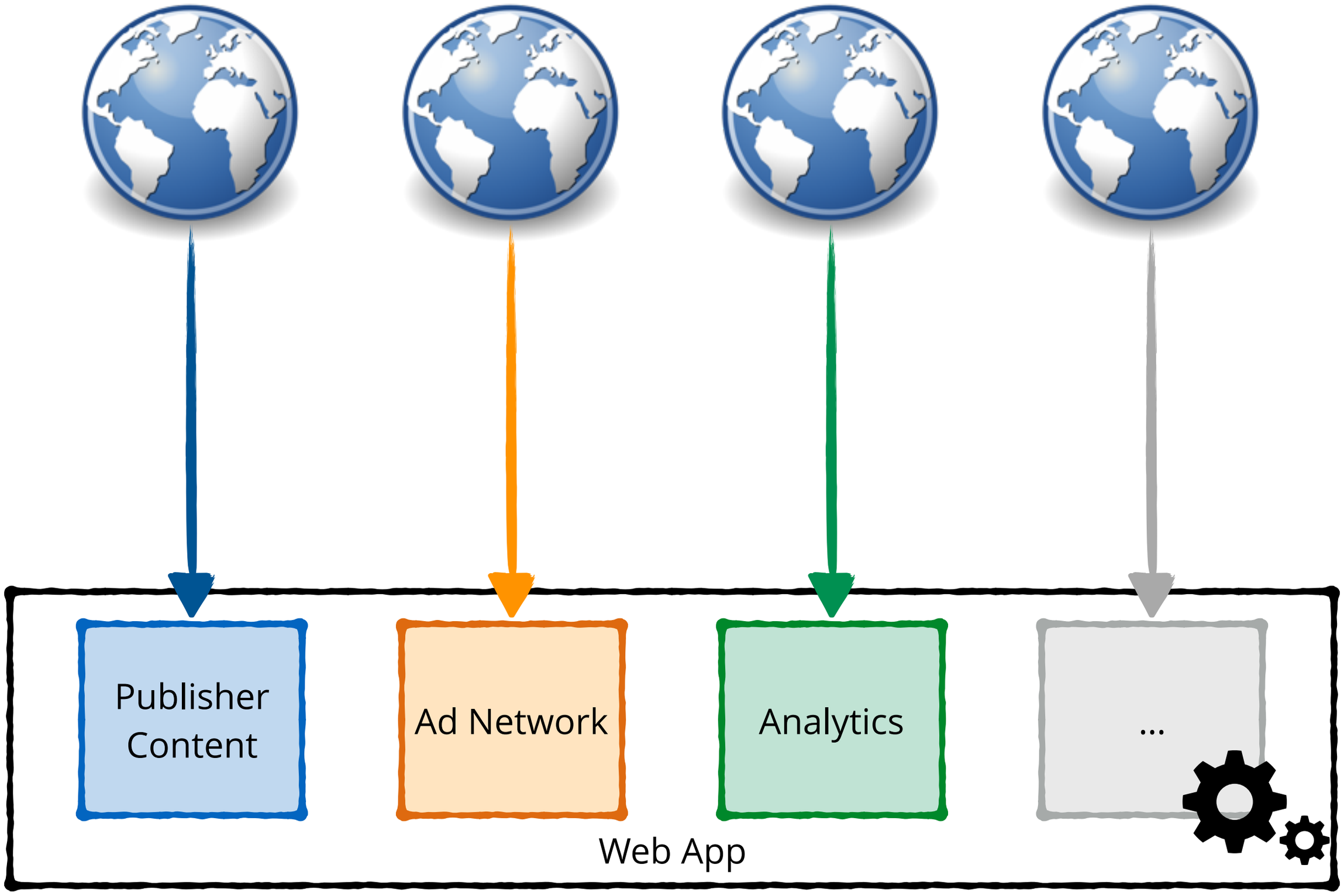
Identifying Extension-based Ad Injection via Fine-grained Web Content Provenance

Sajjad Arshad, Amin Kharraz, and William Robertson
Northeastern University

RAID 2016


Publisher Content
Ad Network
Analytics
...
Web App

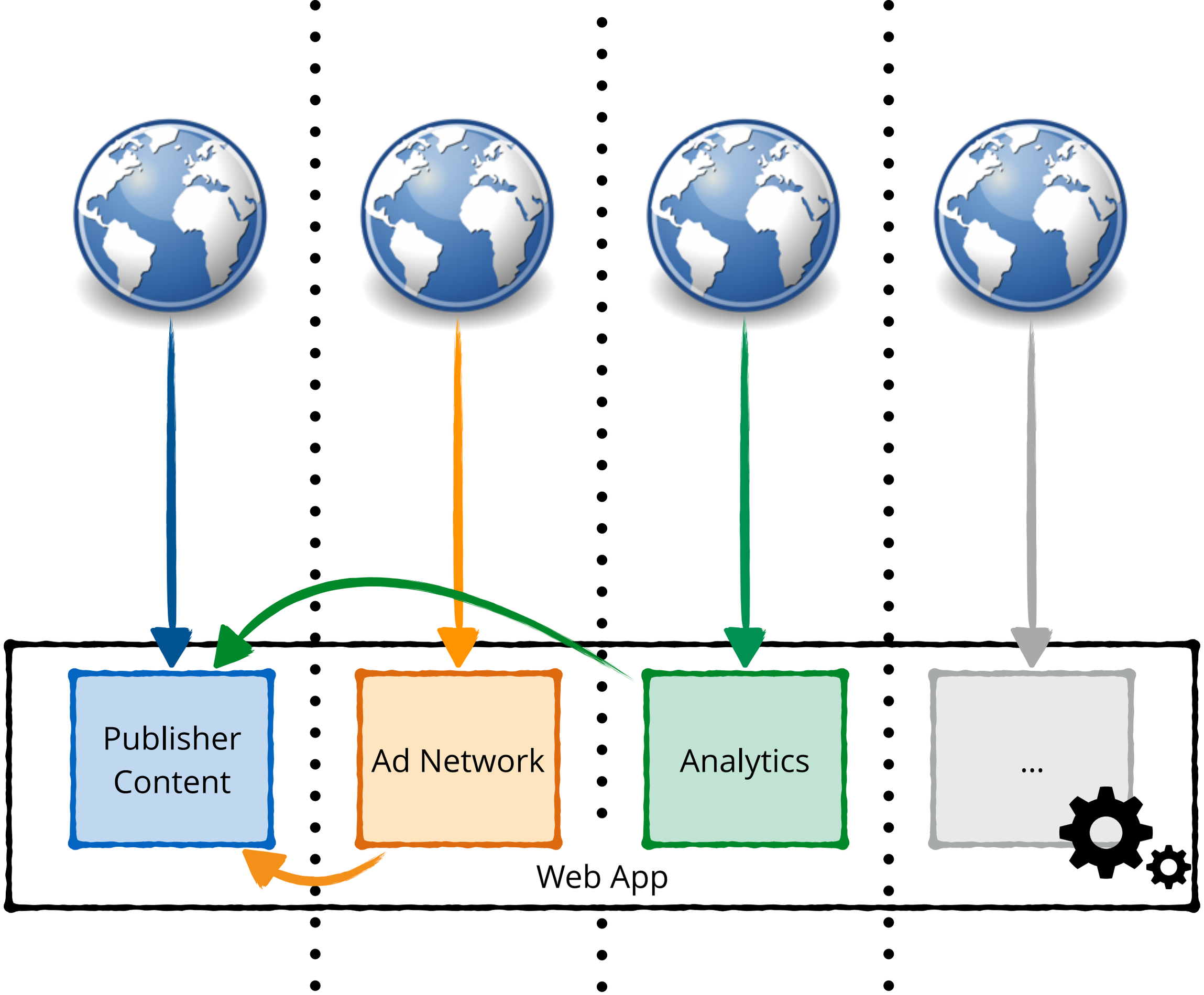
Publisher Content

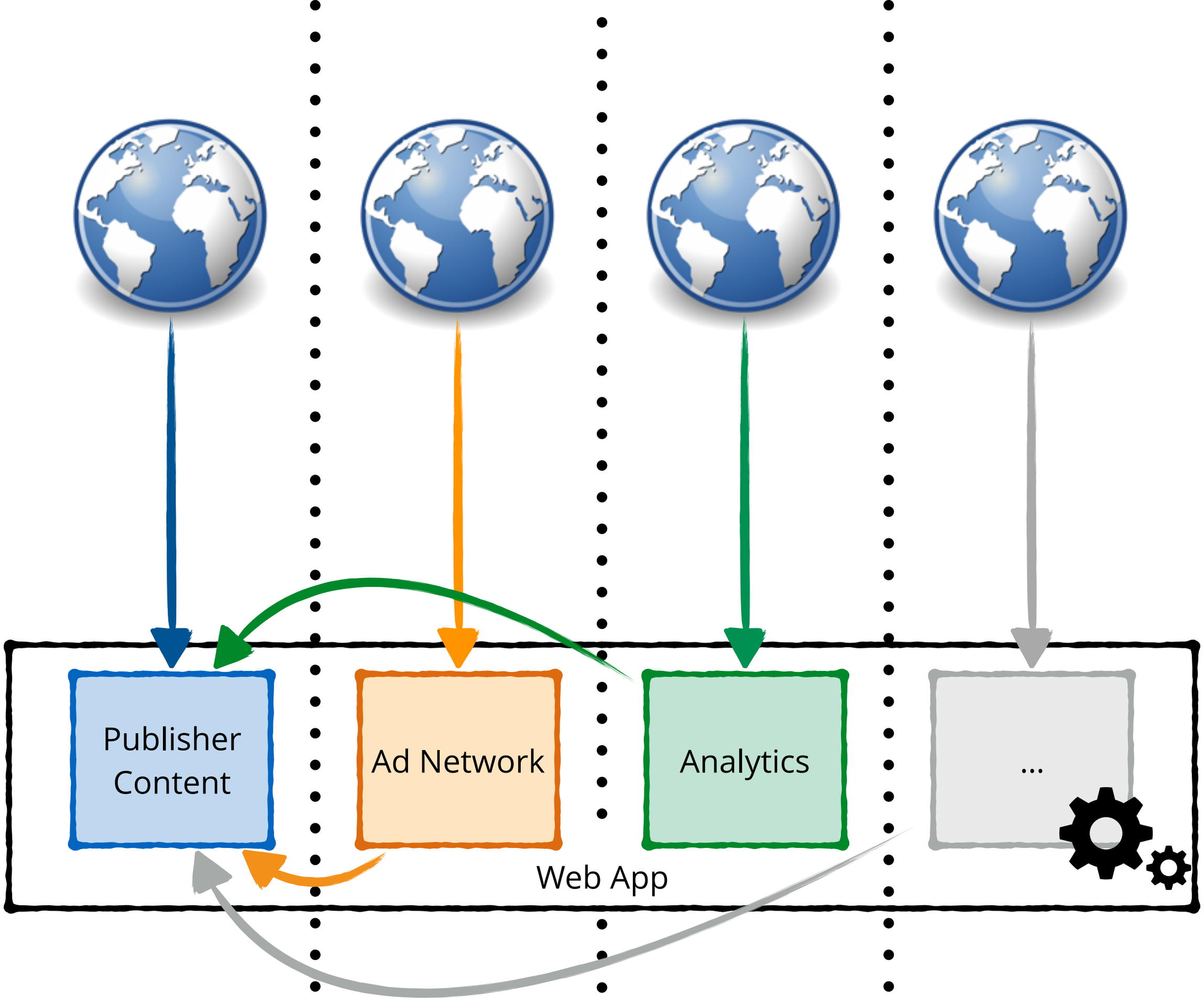
Ad Network

Analytics

...

Web App


Publisher Content
Ad Network
Analytics
...
Web App


Publisher Content
Ad Network
Analytics
...
Web App

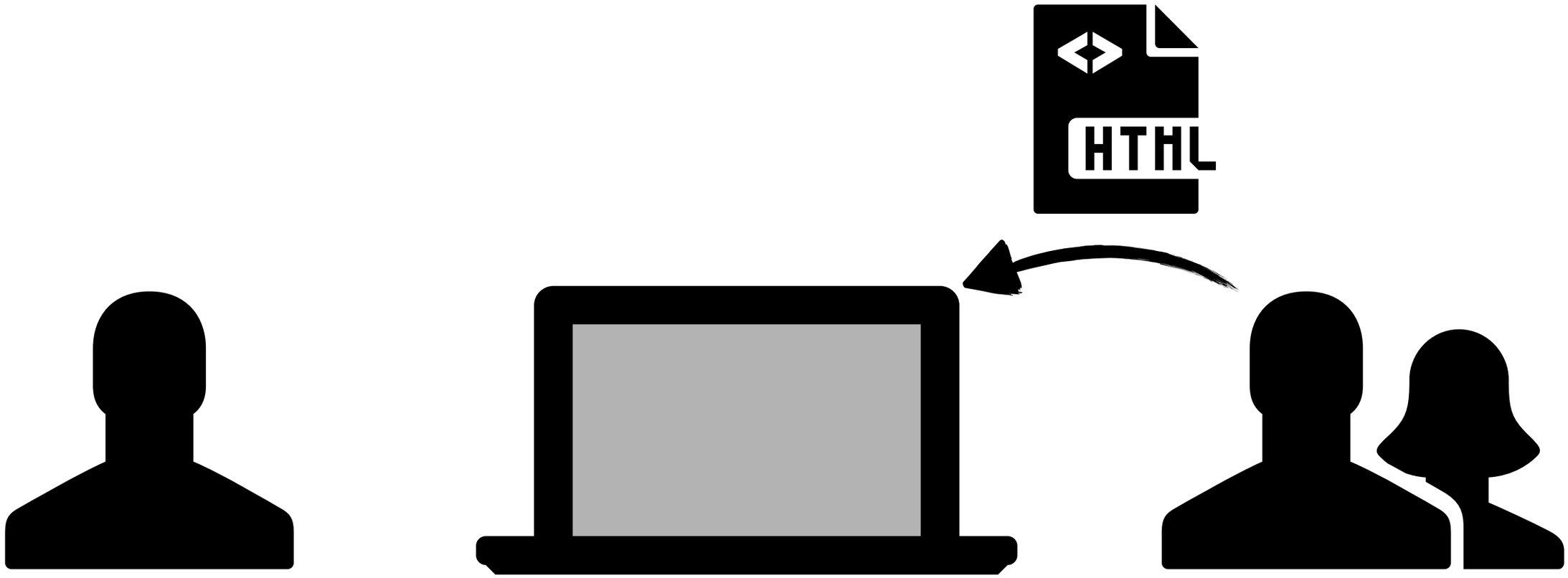
Publisher Content

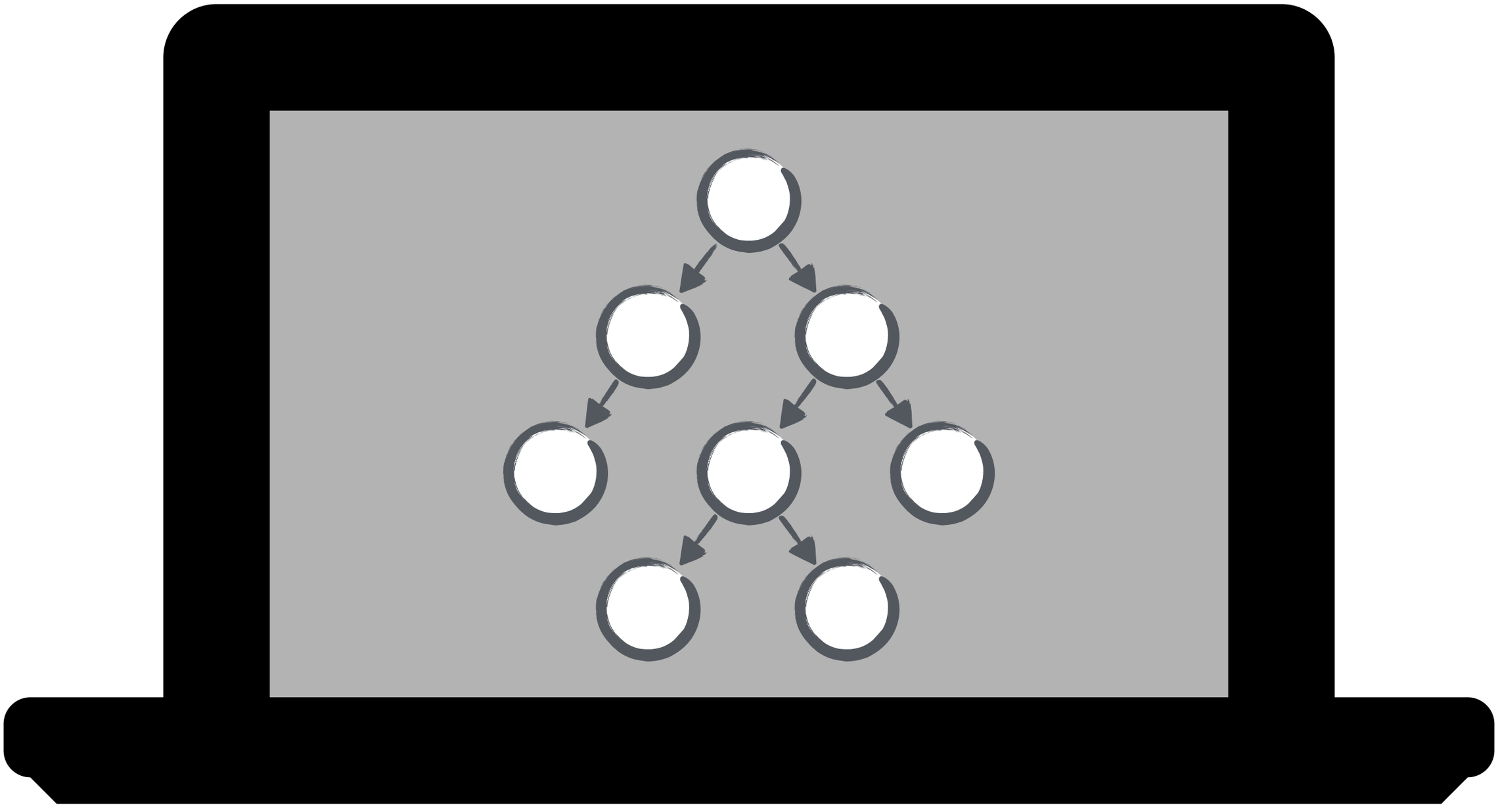
Ad Network

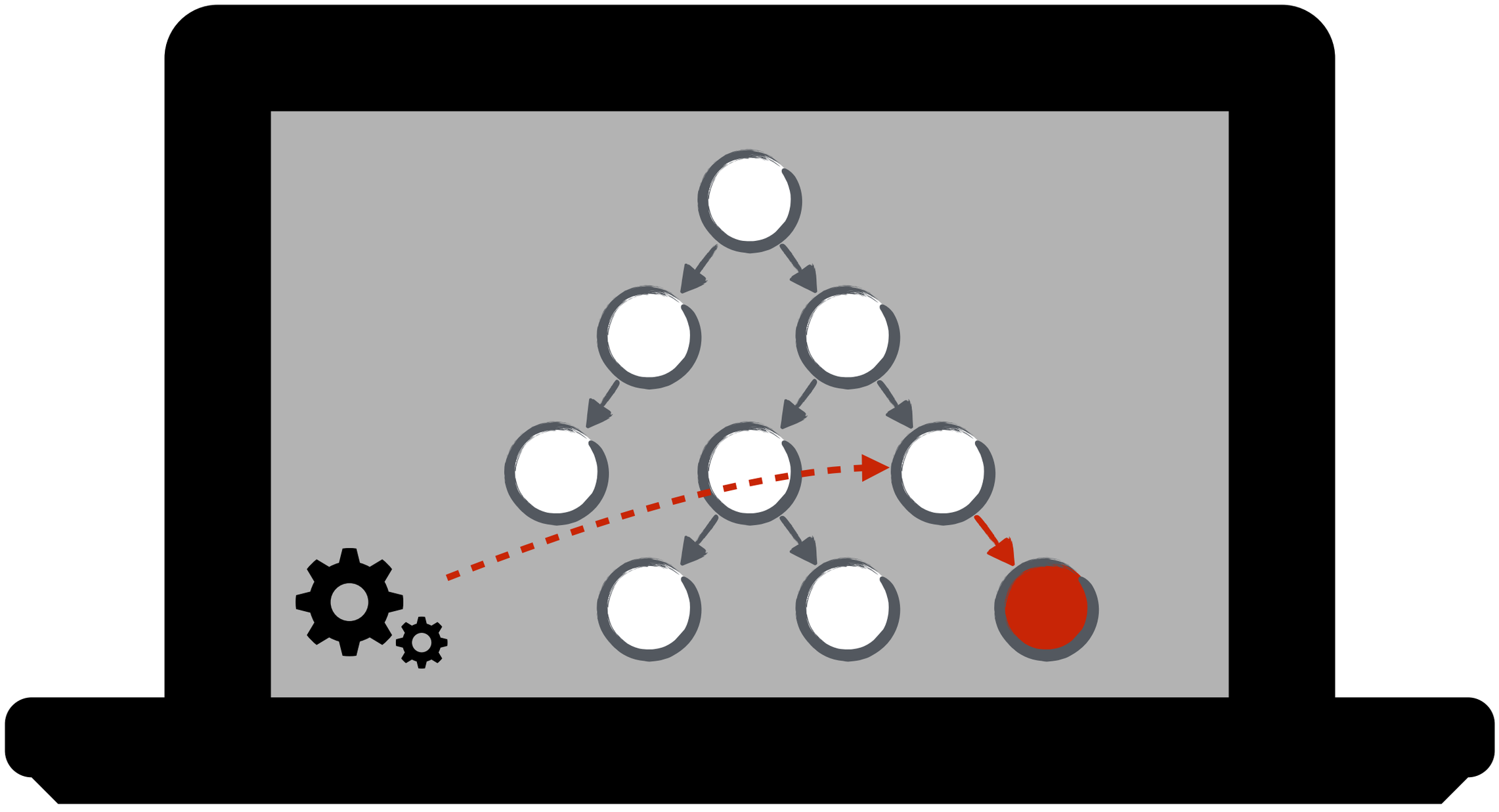
Analytics

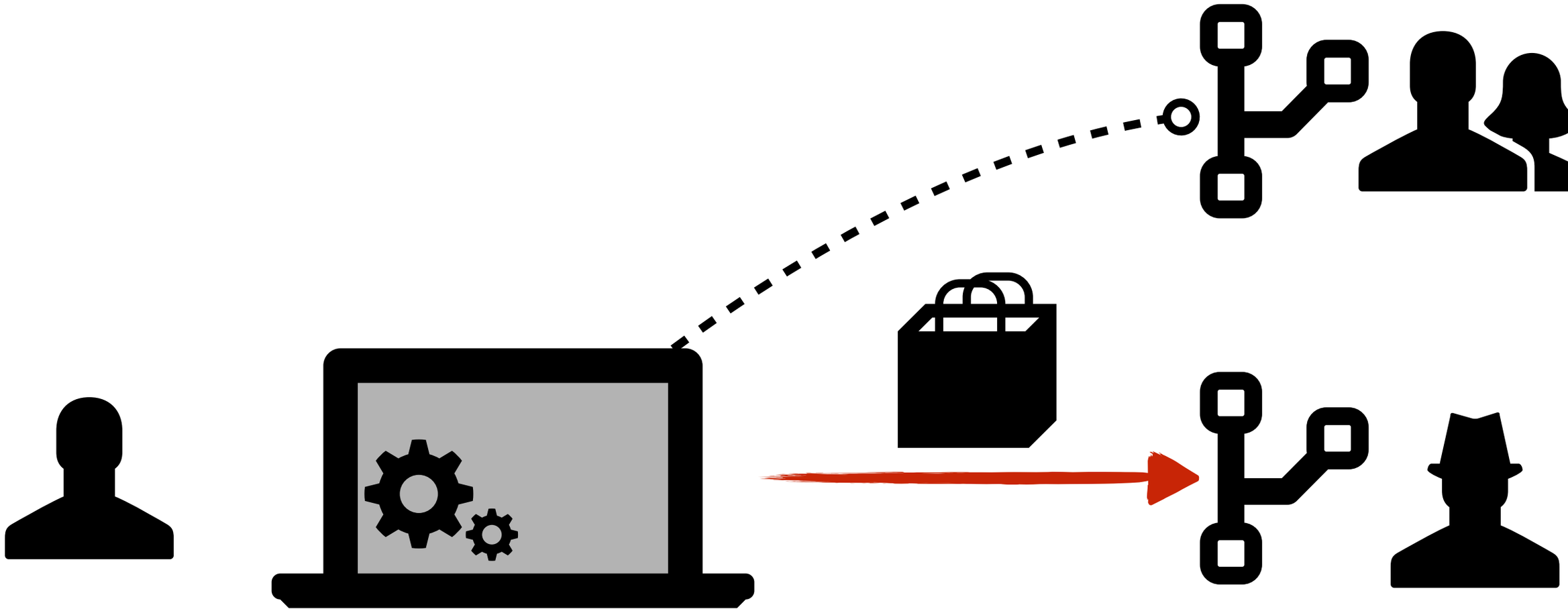
...

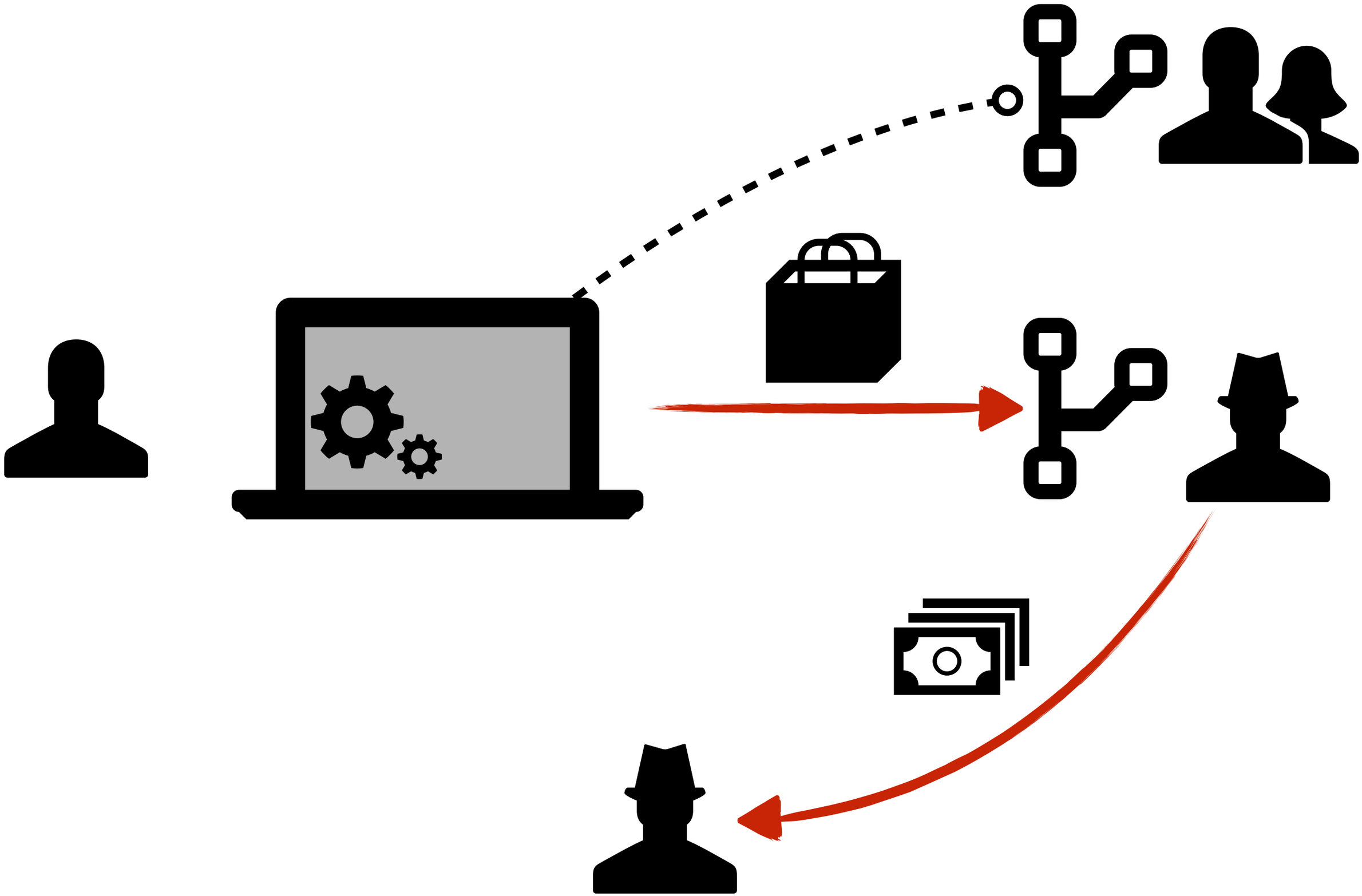
Web App


HTML









Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more - Windows Internet Explorer

http://www.amazon.com/

Bing

Favorites | Amazon.com: Online Shopping for Electronics, Appare... | Page | Safety | Tools

amazon Try Prime

Your Amazon.com | Today's Deals | Gift Cards | Sell | Help

All-New kindle paperwhite
From $119 › Pre-order now

You need to update your version of media player. Update now.

Shop by Department | Search All | Go | Hello. Sign in Your Account | Try Prime | 0 Cart | Wish List

All-New
kindle paperwhite
The best device for reading, period.

From $119 › Pre-order now

Done | Internet | 100%

Start | Fiddler - HTTP Debuggin... | Amazon.com: Online ... | « 2:53 PM

Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more - Windows Internet Explorer

http://www.amazon.com/

Bing

Favorites | Amazon.com: Online Shopping for Electronics, Appare... | Page | Safety | Tools

amazon Try Prime

Your Amazon.com | Today's Deals | Gift Cards | Sell | Help

All-New kindle paperwhite
From $119 › Pre-order now

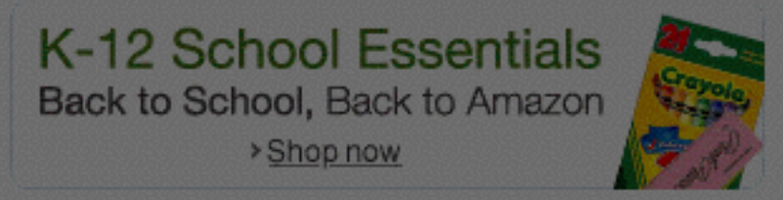
You need to update your version of media player. Update now.

Shop by Department | Search All | Go | Hello. Sign in Your Account | Try Prime | 0 Cart | Wish List

All-New
kindle paperwhite
The best device for reading, period.

From $119 › Pre-order now

Amazon Fashion
New Dresses
The latest wear-everywhere styles—from prints to fit-and-flare.



Done | Internet | 100%

Start | Fiddler - HTTP Debuggin... | Amazon.com: Online ... | « 2:53 PM

Detecting Ad Injection

- Centralized dynamic analysis at extension distribution points
- Dynamic analysis can be effective, but is also prone to the usual caveats
- Third-party content injection or modification by extensions is quite common
- Can be difficult for browser vendors to delineate between wanted and unwanted behavior

Users are best positioned to make this judgment

OriginTracer

OriginTracer

OriginTracer adds fine-grained content provenance tracking to the web browser

- Provenance tracked at level of individual DOM elements
- Indicates origins contributing to content injection and modification
- Trustworthy communication of this information to the user

PROVENANCE LABELS

- Labels are generalizations of web origins

$$L = \langle S, I, P, X \rangle$$

$$S = \{\text{scheme}\} \bigcup \{\text{"extension"}\}$$

$$I = \{\text{host}\} \bigcup \{\text{extension-id}\}$$

$$P = \{\text{port}\} \bigcup \{\text{null}\}$$

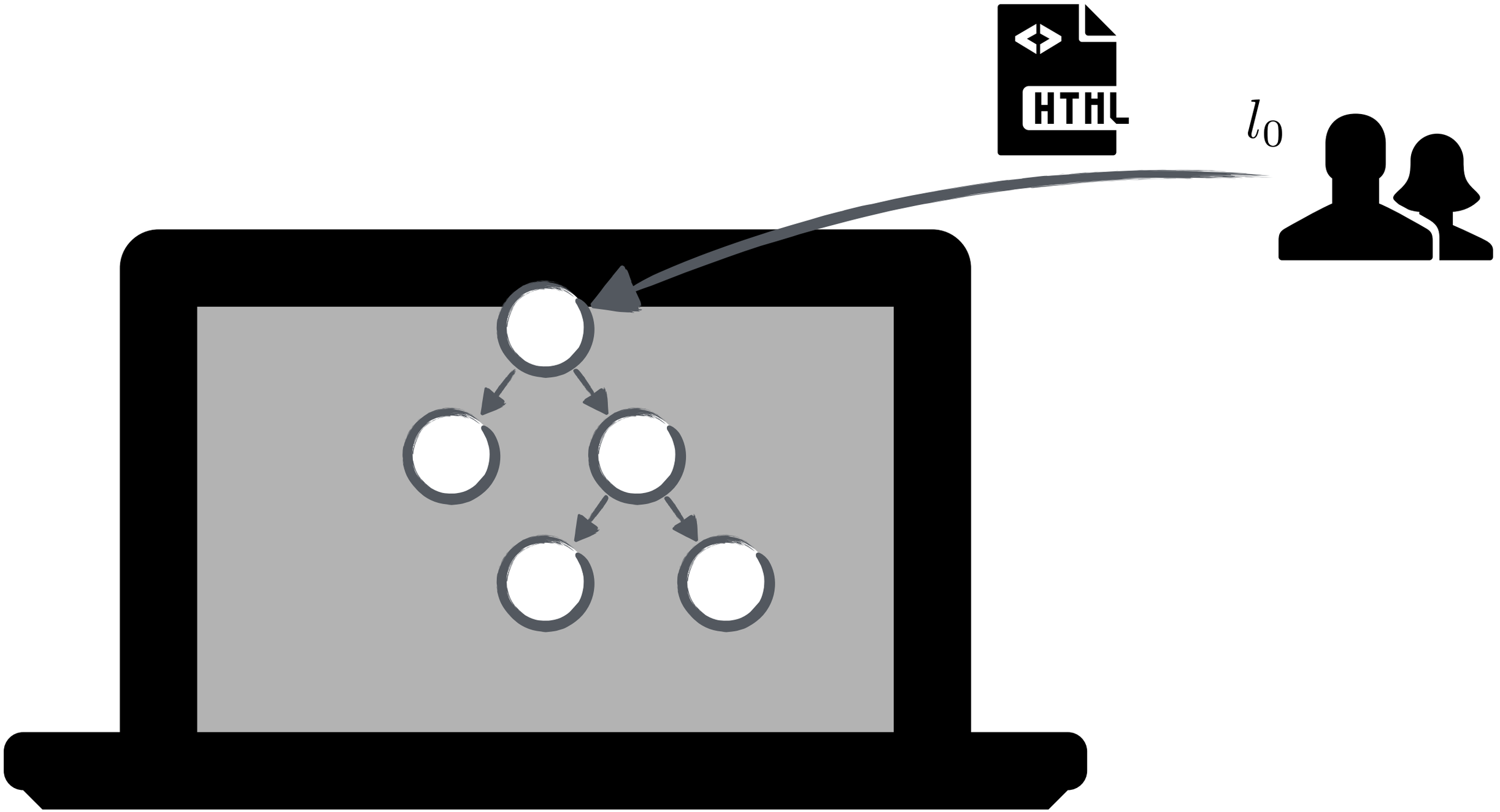
$$X = \{0, 1, 2, \ldots\}$$

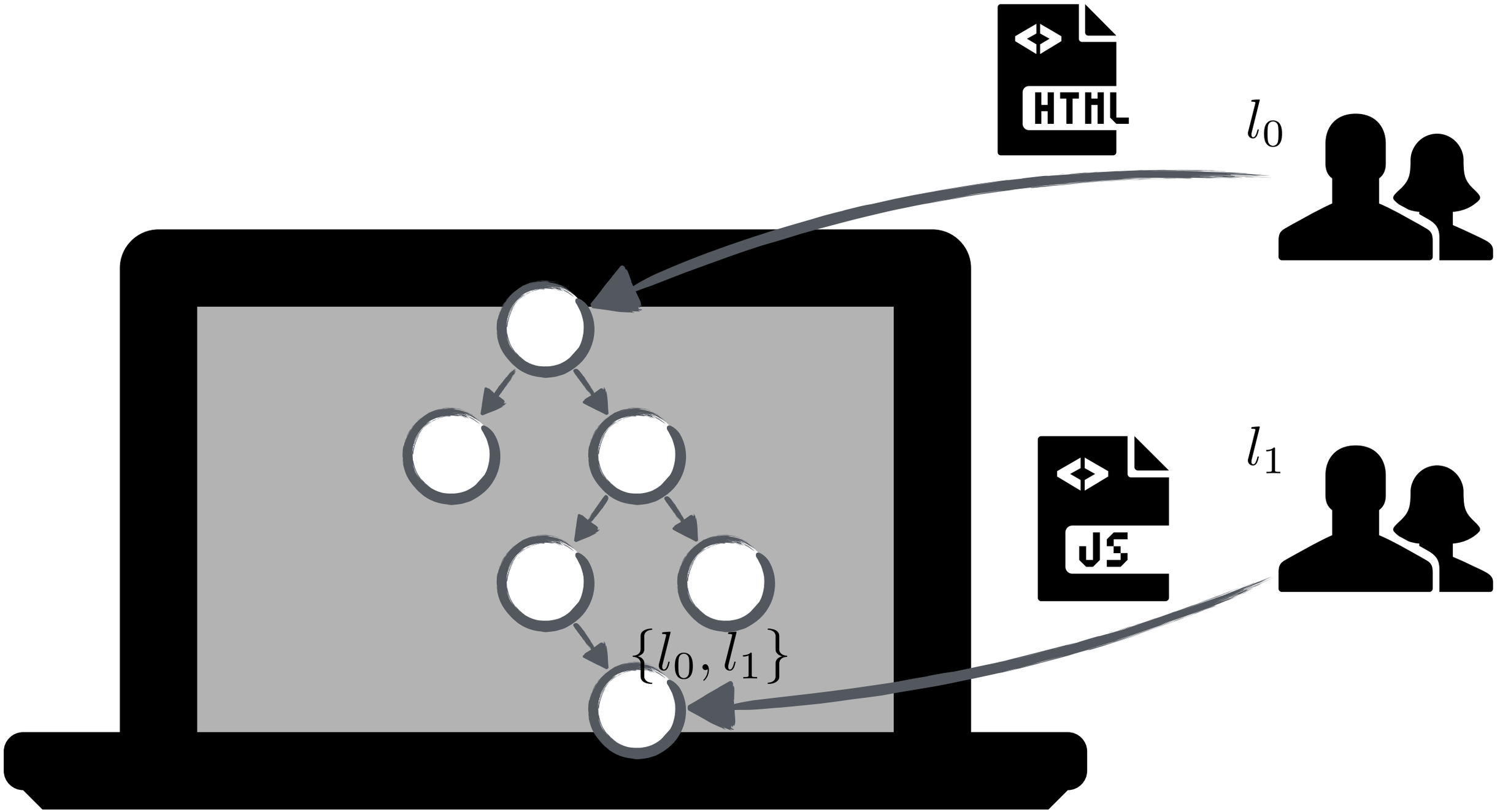
Label Propagation

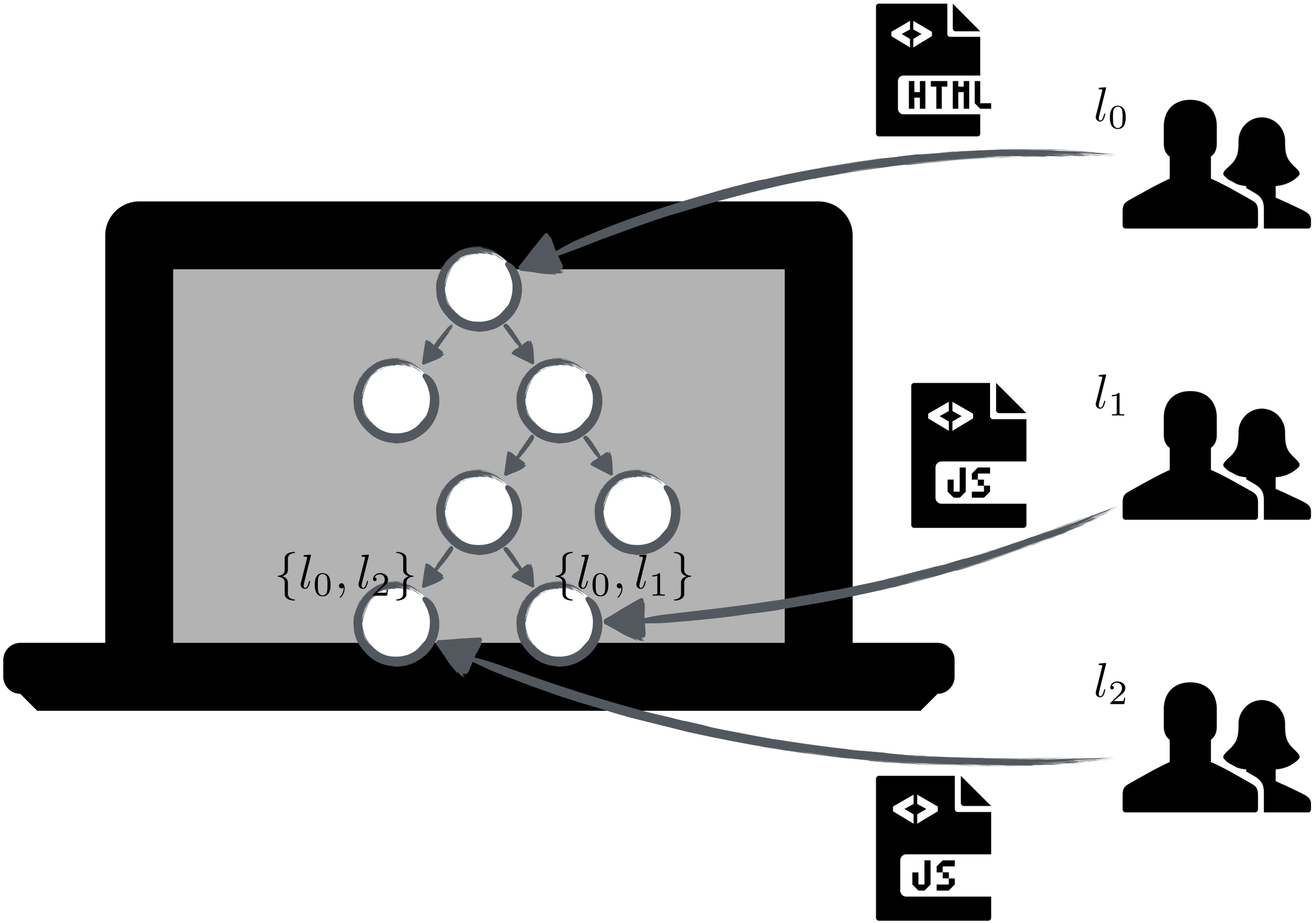
- Static content is assigned the publisher's label $\{l_0\}$
- Dynamic content due to external scripts
  - New external scripts are assigned a label $\{l_i\}$
    - Injected or modified content is labeled $\{l_0, l_i\}$
- Extension content
  - Initialized with unique label as for external scripts
  - But, injected or modified content labels omit the publisher's label ($\{l_0, l_j\}$)

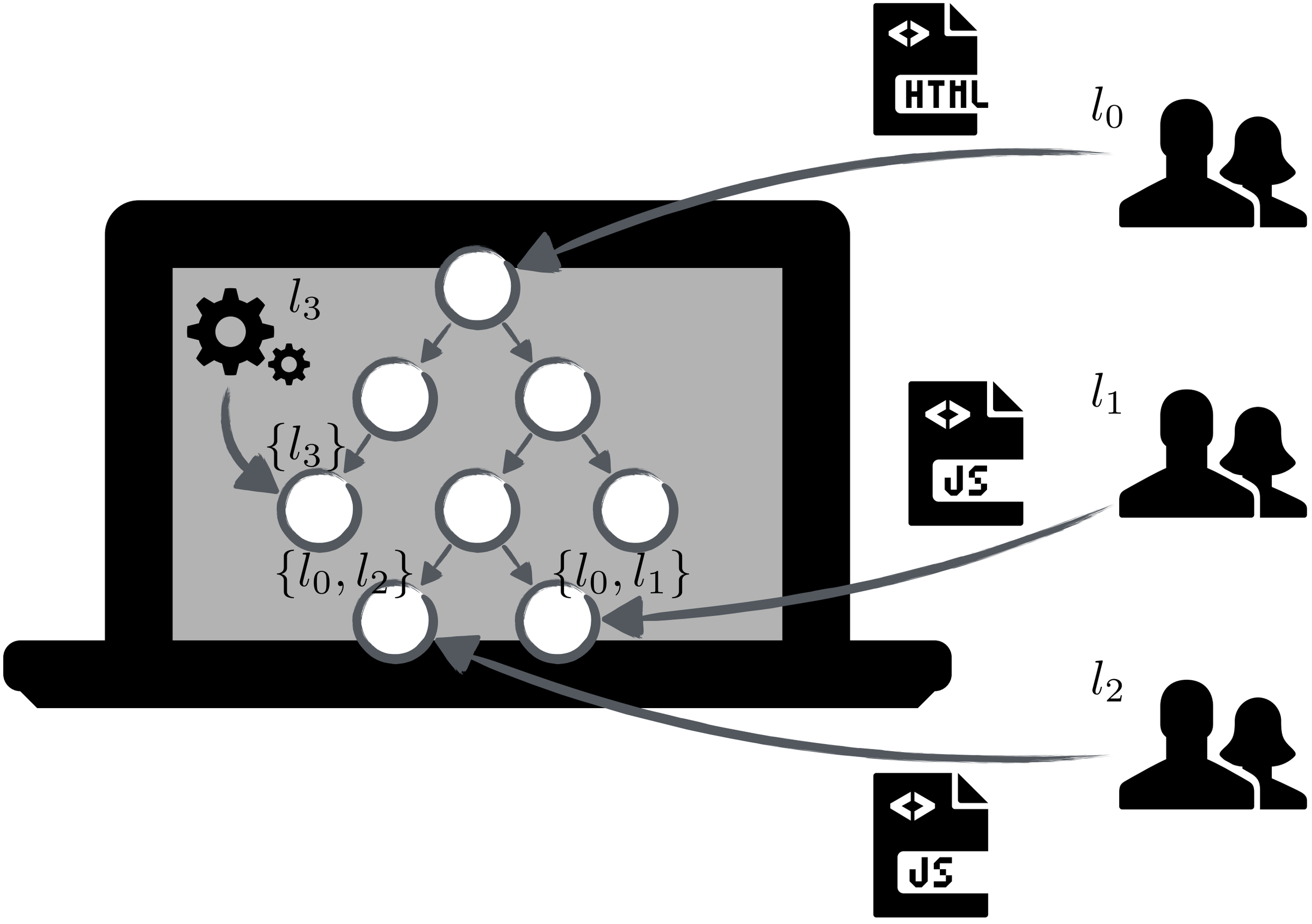
Label Propagation

- Static content is assigned the publisher's label $\{l_0\}$
- Dynamic content due to external scripts
  - New external scripts are assigned a label $\{l_i\}$
    - Injected or modified content is labeled $\{l_0, l_i\}$
- Extension content
  - Initialized with unique label as for external scripts
  - But, injected or modified content labels omit the publisher's label ($\{\cancel{l_0}, l_j\}$)


HTML
l_0


HTML
l_0
l_1
JS
$\{l_0, l_1\}$


HTML
l_0
l_1
JS
$\{l_0, l_2\}$
$\{l_0, l_1\}$
l_2
JS


HTML
l_0
l_3
$\{l_3\}$
$\{l_0, l_2\}$
$\{l_0, l_1\}$
JS
l_1
l_2
JS

Provenance Indicators

- Provenance must be communicated to the user in a trustworthy way
- What is the best way to communicate provenance?
  - Full provenance label sets likely to be difficult to comprehend
  - For extensions, we chose to use the extension title
  - For arbitrary content, origins are easy but not ideal

Implementation

- Modifications to Chromium browser
  - ~900 SLoC (C++), several lines of JavaScript
  - Mediates DOM APIs for node creation and modification
  - Mediates node insertion through document writes
  - Callbacks registered for events and timers

PC Build of the Month

's Deals
new deals
ry day

By X-notifier

Injected by X-Notifier Extension

Evaluation

Evaluation Questions

(1) How susceptible are users to injected content?

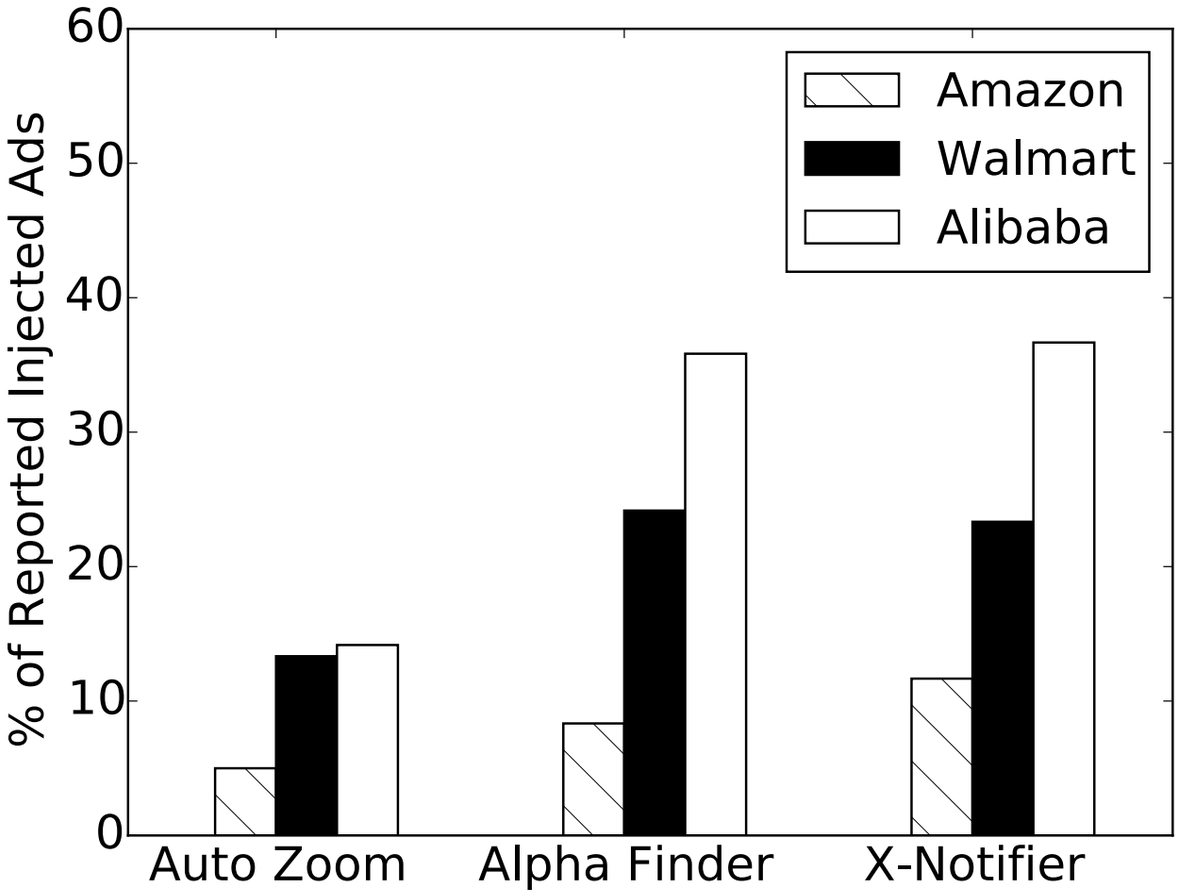
(2) Do provenance indicators reduce clicks on extension-injected content?

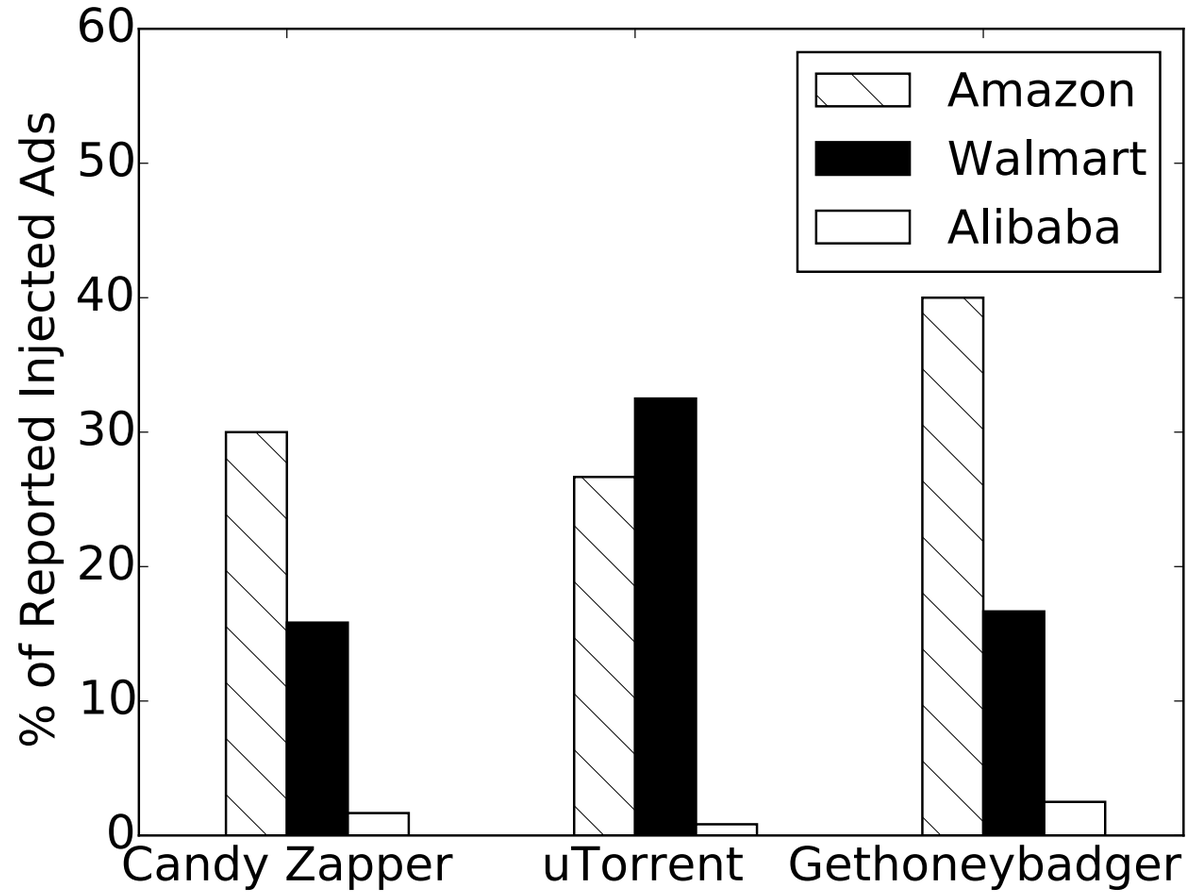
(3) Would users adopt a provenance tracking system?

(4) Does provenance tracking degrade browser performance and user experience?

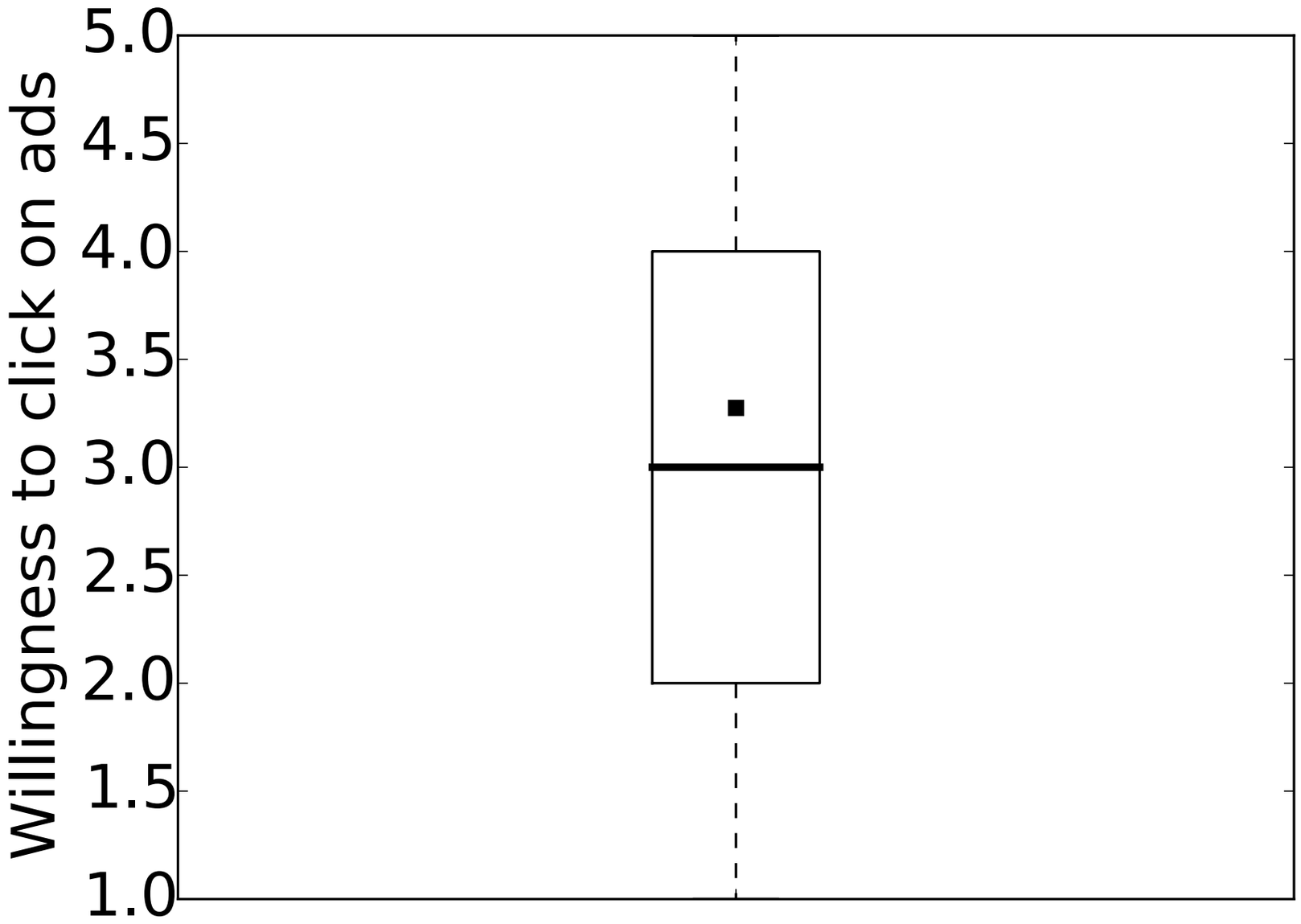
User Study Setup

- Study population: 80 students of varying technical sophistication
- Participants exposed to six Chromium instances (unmodified and modified), each with an ad-injecting extension installed
  - Auto Zoom, Alpha Finder, X-Notifier, Candy Zapper, uTorrent, Gethoneybadger
- Participants were asked to visit three retail websites
  - Amazon, Walmart, Alibaba

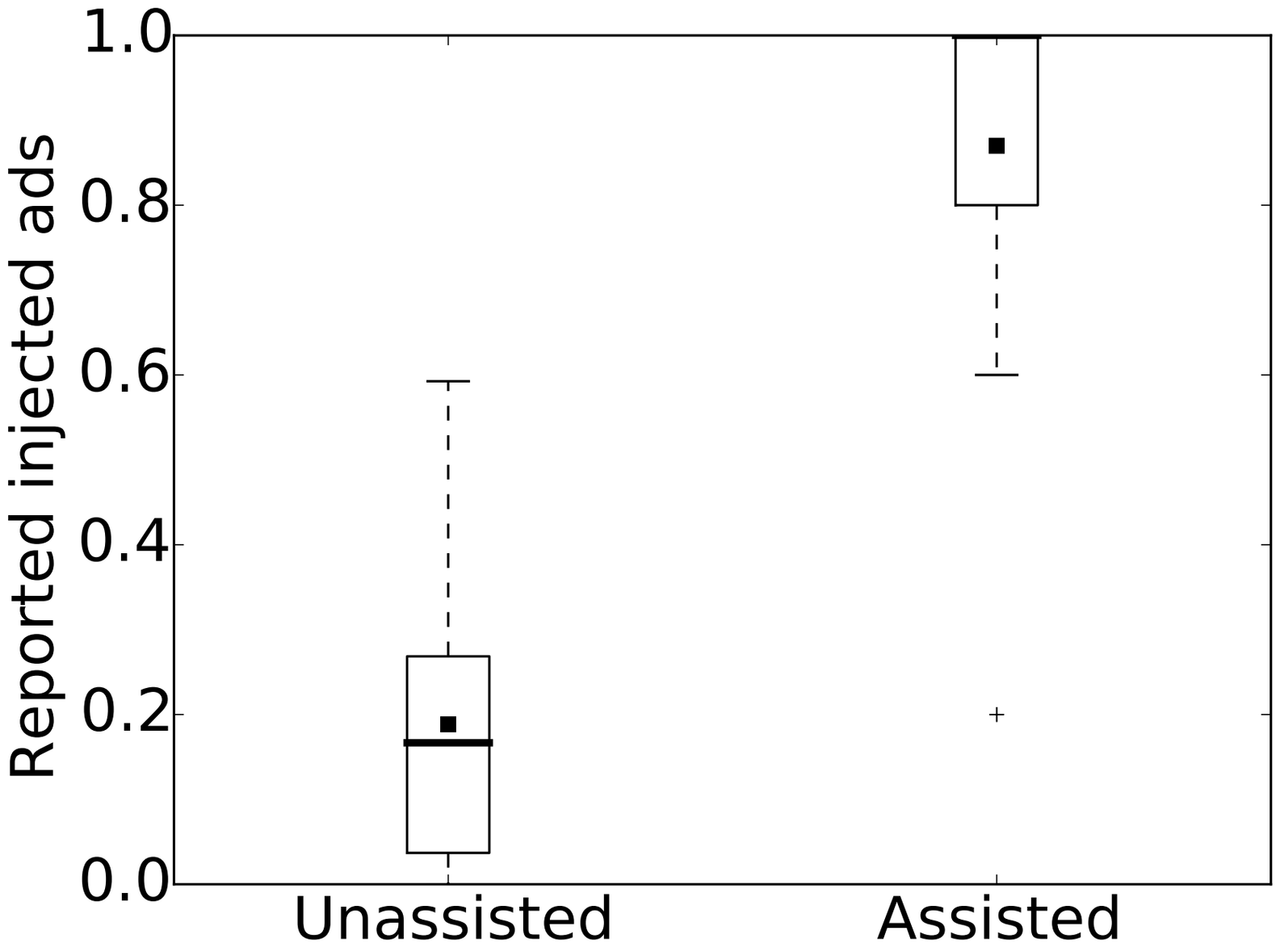




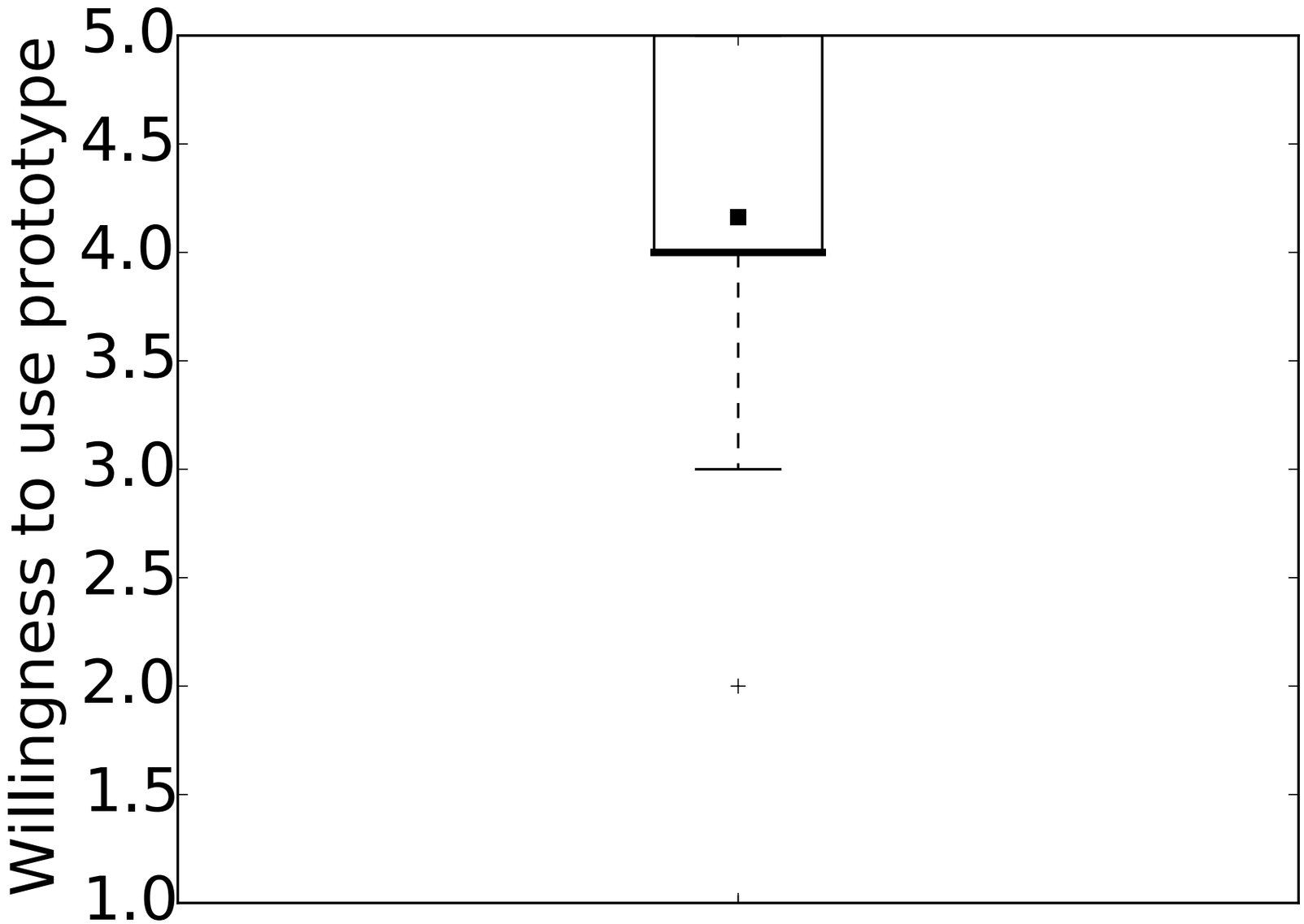
Are users able to correctly recognize injected advertisements?



Are users generally willing to click on the advertisements presented to them?



Do content provenance indicators assist users in recognizing injected advertisements?



Would users be willing to adopt a provenance tracking system to identify injected advertisements?

Reliability

- Separate user study on 13 students of varying technical background
- Asked participants to browse the web for several hours using the OriginTracer prototype
- Asked users to report errors
  - Type I: browser crash, page doesn't load, etc.
  - Type II: abnormal load time, page appearance not as expected
- Out of almost 2K URLs, two Type I and 27 Type II errors were reported

Performance Overhead

- Configured an unmodified Chromium and OriginTracer instance to visit the Alexa Top 1K
  - Broad spectrum of static and dynamic content on most-used websites
  - Browsers configured with five benign extensions
- Average 10.5% browsing latency overhead
- No impact on browser start-up time

Future Work

- Usability of provenance indicators should be considered an initial attempt
  - Many points in the design space we did not explore
  - Extending provenance for other applications
    - Surfacing fine-grained provenance is potentially highly useful
    - e.g., inferring remote data flows between origins
    - e.g., fine-grained policy enforcement

Conclusions

- Some forms of questionable behavior on the web are best judged by the user
- OriginTracer tracks web content provenance in a fine-grained way, allowing users to make similarly fine-grained trust decisions
- Evaluation shows that provenance tracking can be performed in an efficient and effective way for modern browsers and web content

THANKS! QUESTIONS?

Sajjad Arshad
<arshad@ccs.neu.edu>