Tracing Information Flows Between Ad Exchanges Using Retargeted Ads

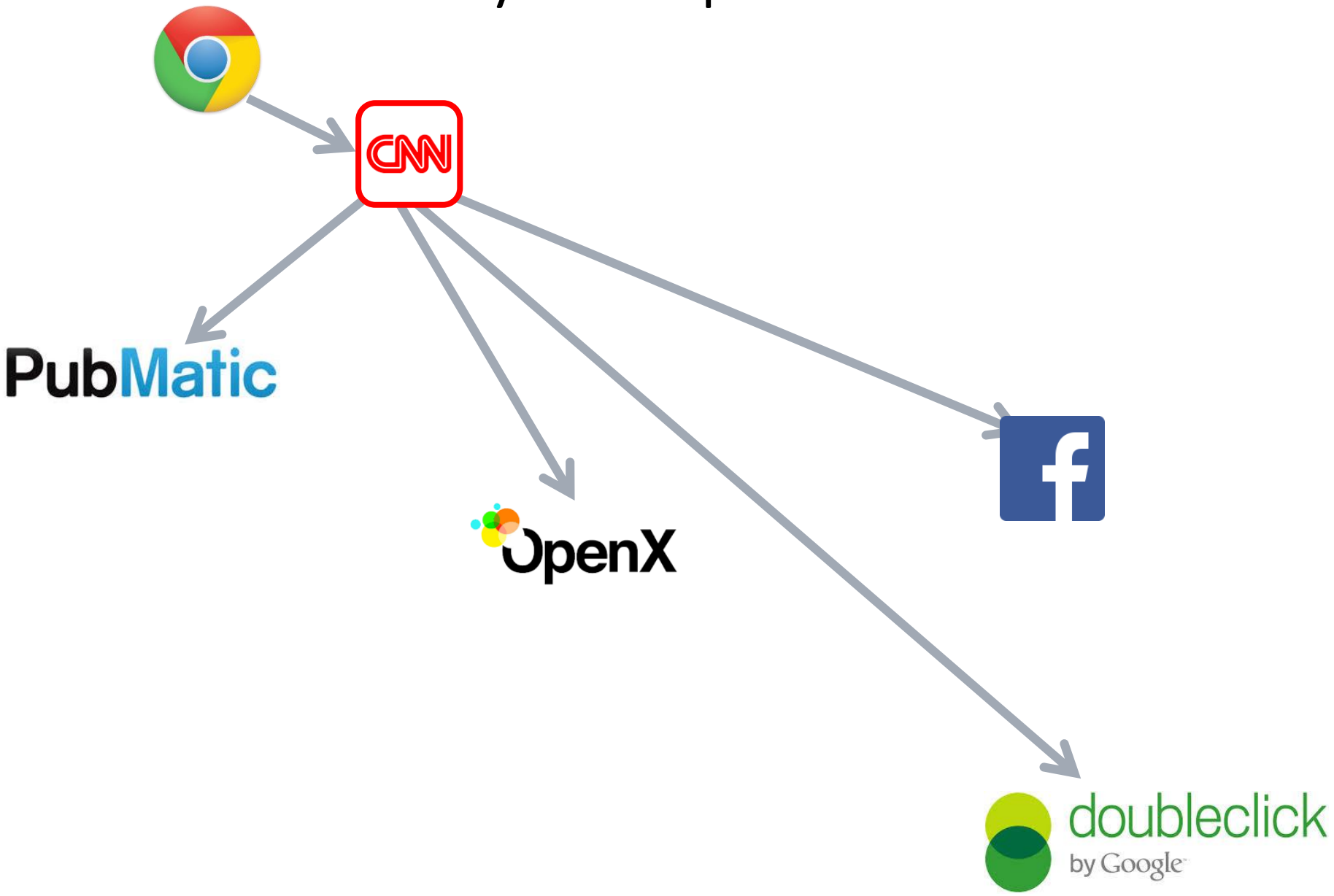
Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, Christo Wilson

Northeastern University

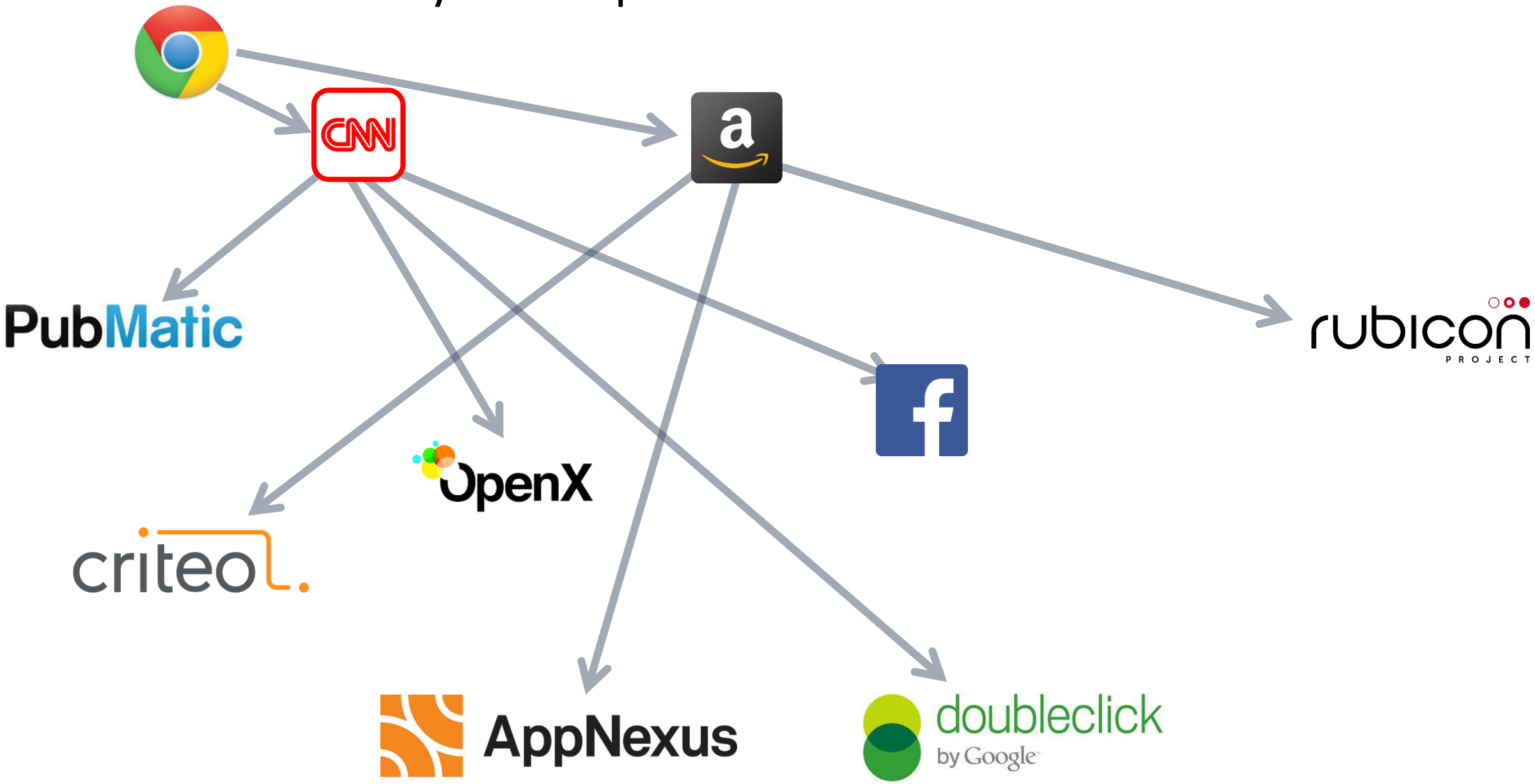
Your Privacy Footprint

Your Privacy Footprint

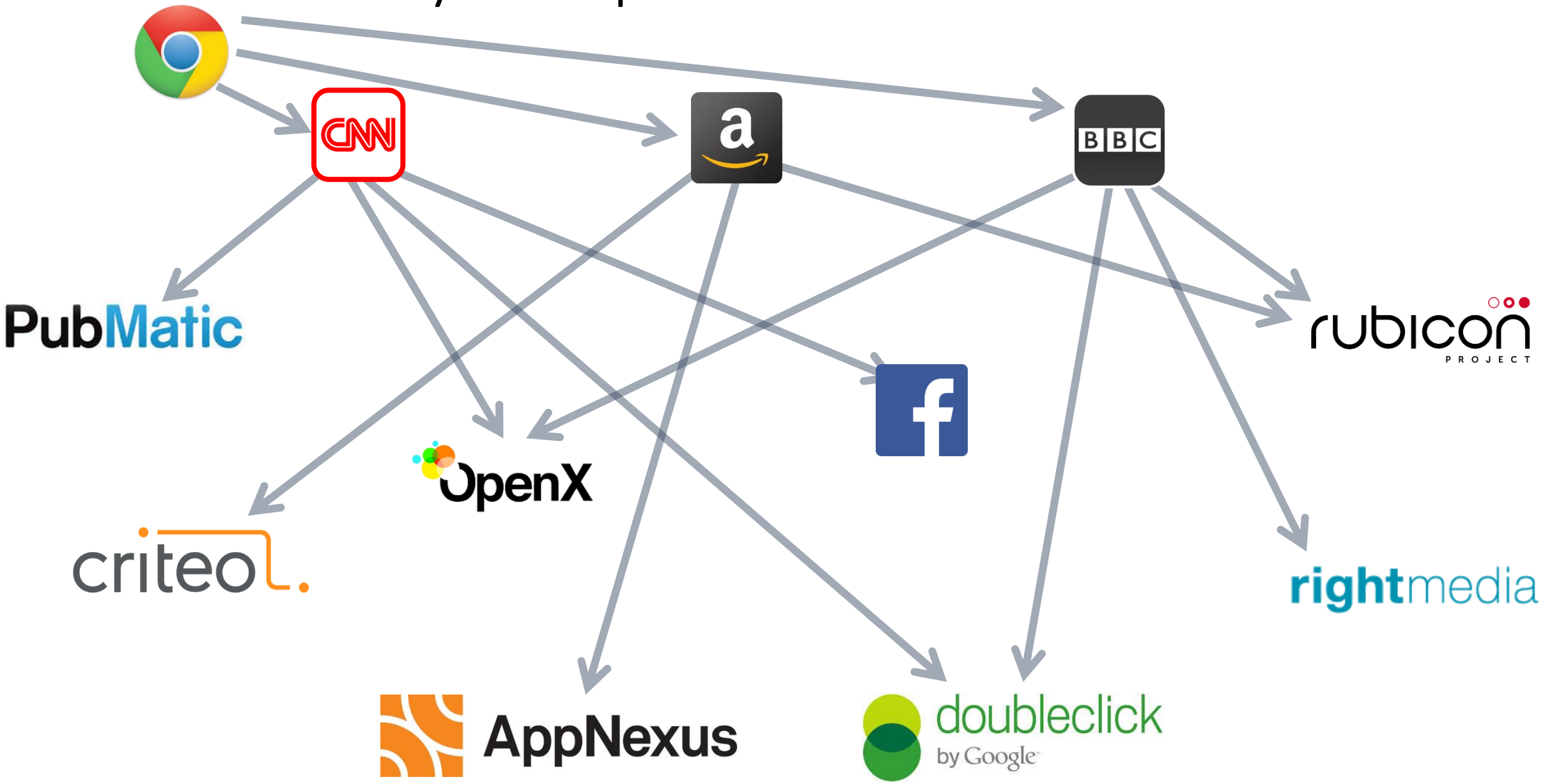
Your Privacy Footprint



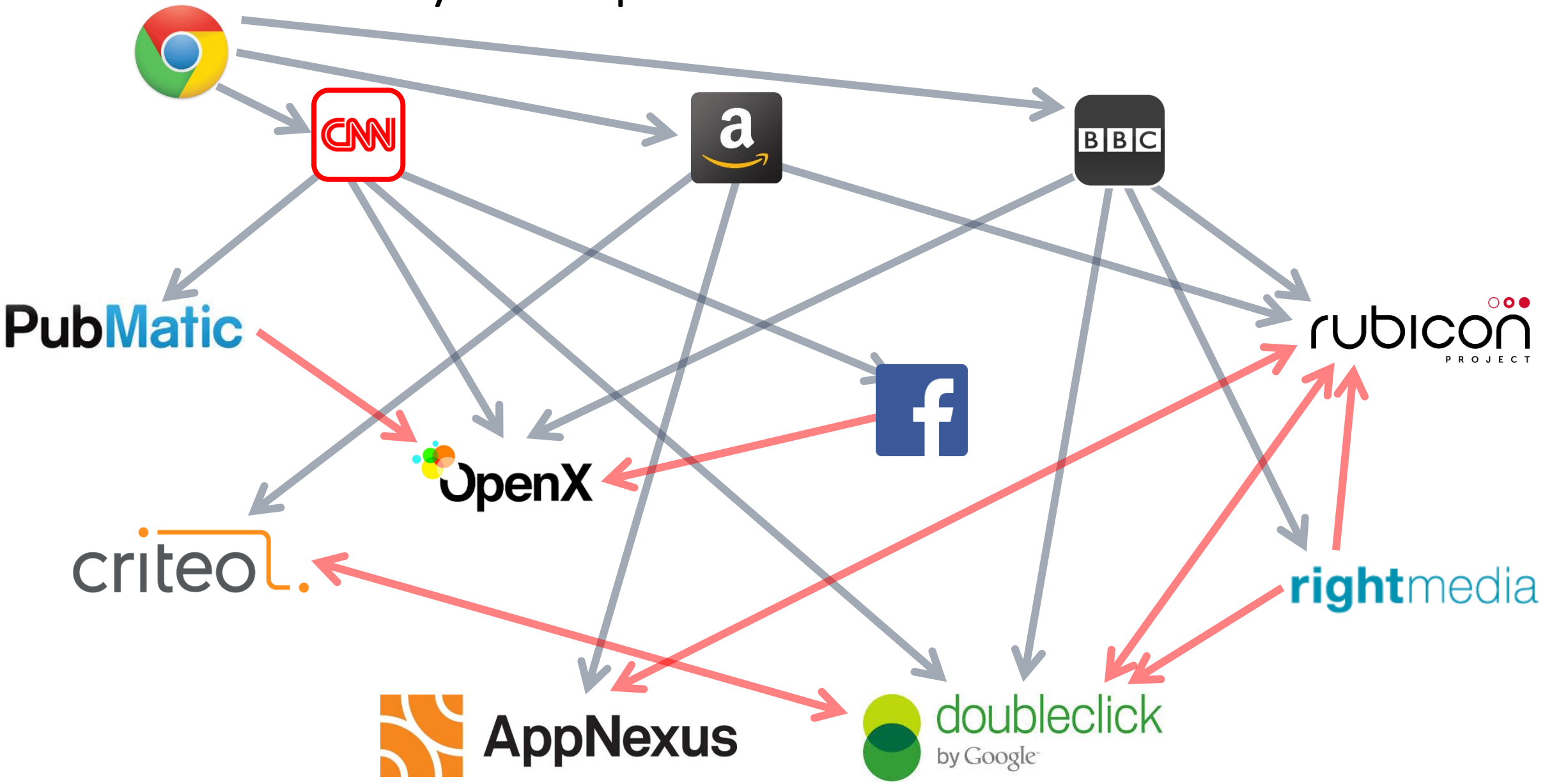
Your Privacy Footprint



Your Privacy Footprint



Your Privacy Footprint



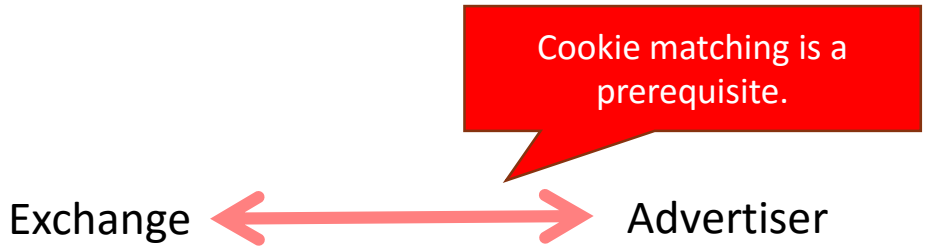
Real Time Bidding

Real Time Bidding

- RTB brings more flexibility in the ad ecosystem.
  - Ad request managed by an Ad Exchange which holds an auction.
  - Advertisers bid on each ad impression.

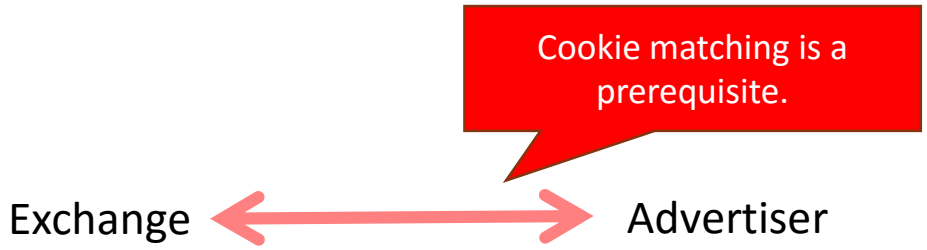
Real Time Bidding

- RTB brings more flexibility in the ad ecosystem.
  - Ad request managed by an Ad Exchange which holds an auction.
  - Advertisers bid on each ad impression.



Real Time Bidding

- RTB brings more flexibility in the ad ecosystem.
  - Ad request managed by an Ad Exchange which holds an auction.
  - Advertisers bid on each ad impression.



- RTB spending to cross $20B by 2017[1].
  - 49% annual growth.
  - Will account for 80% of US Display Ad spending by 2022.

[1] http://www.prnewswire.com/news-releases/new-idc-study-shows-real-time-bidding-rtb-display-ad-spend-to-grow-worldwide-to-208-billion-by-2017-228061051.html

User

Publisher

Ad Exchange

Advertisers

User

Publisher

Ad Exchange

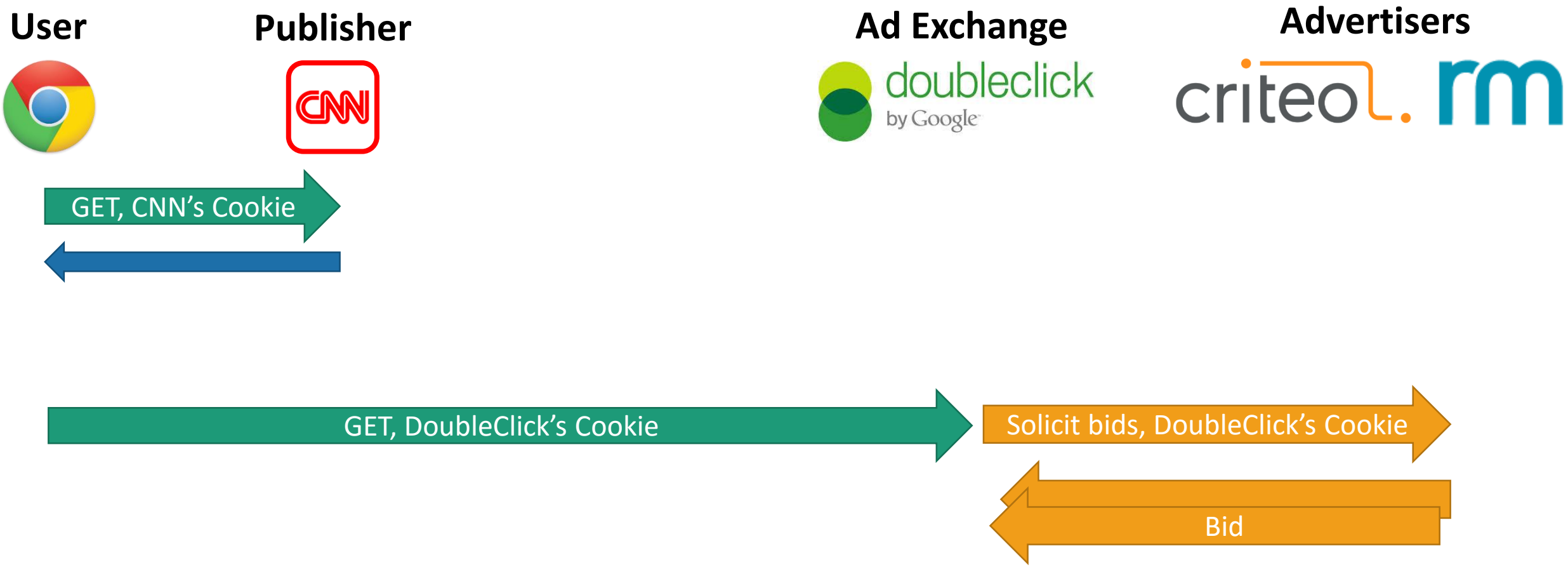
Advertisers

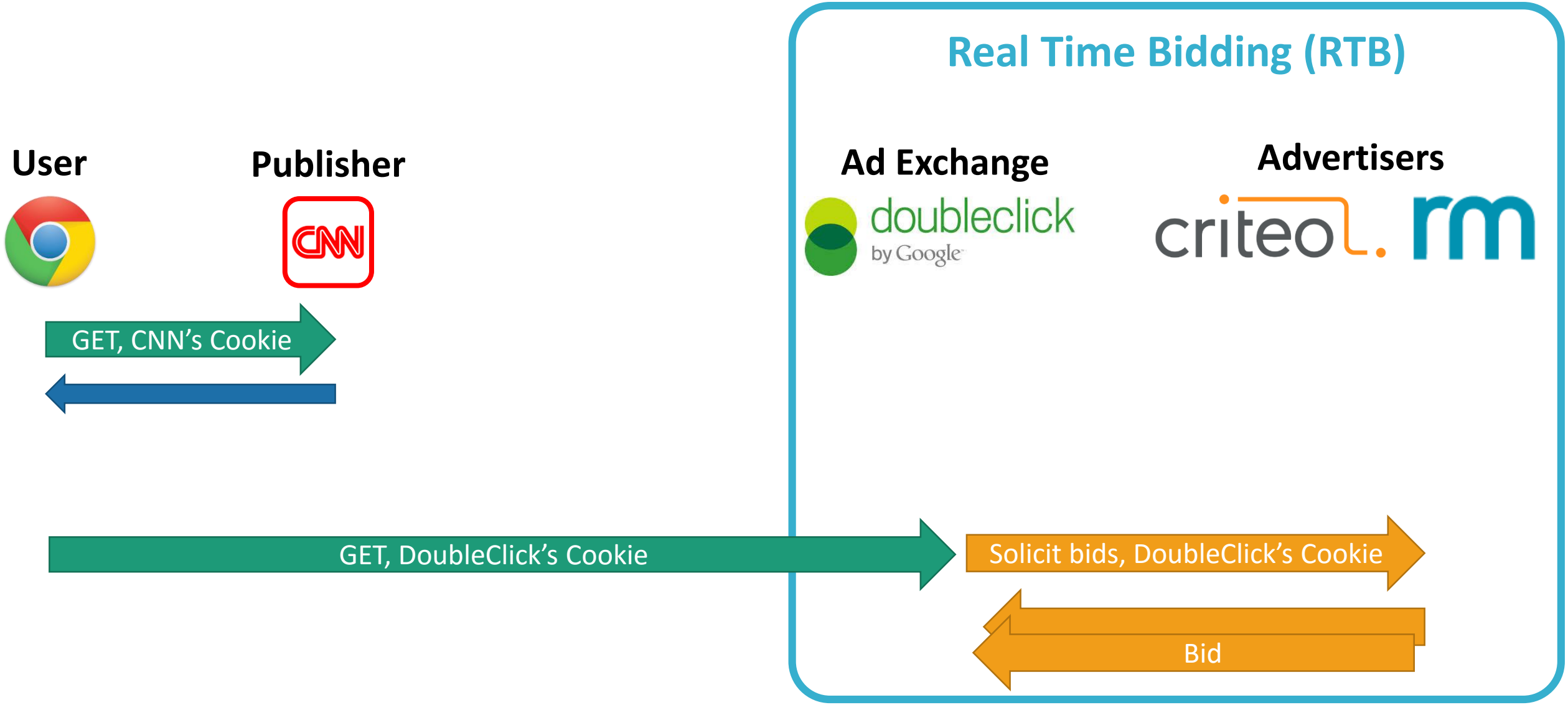
User

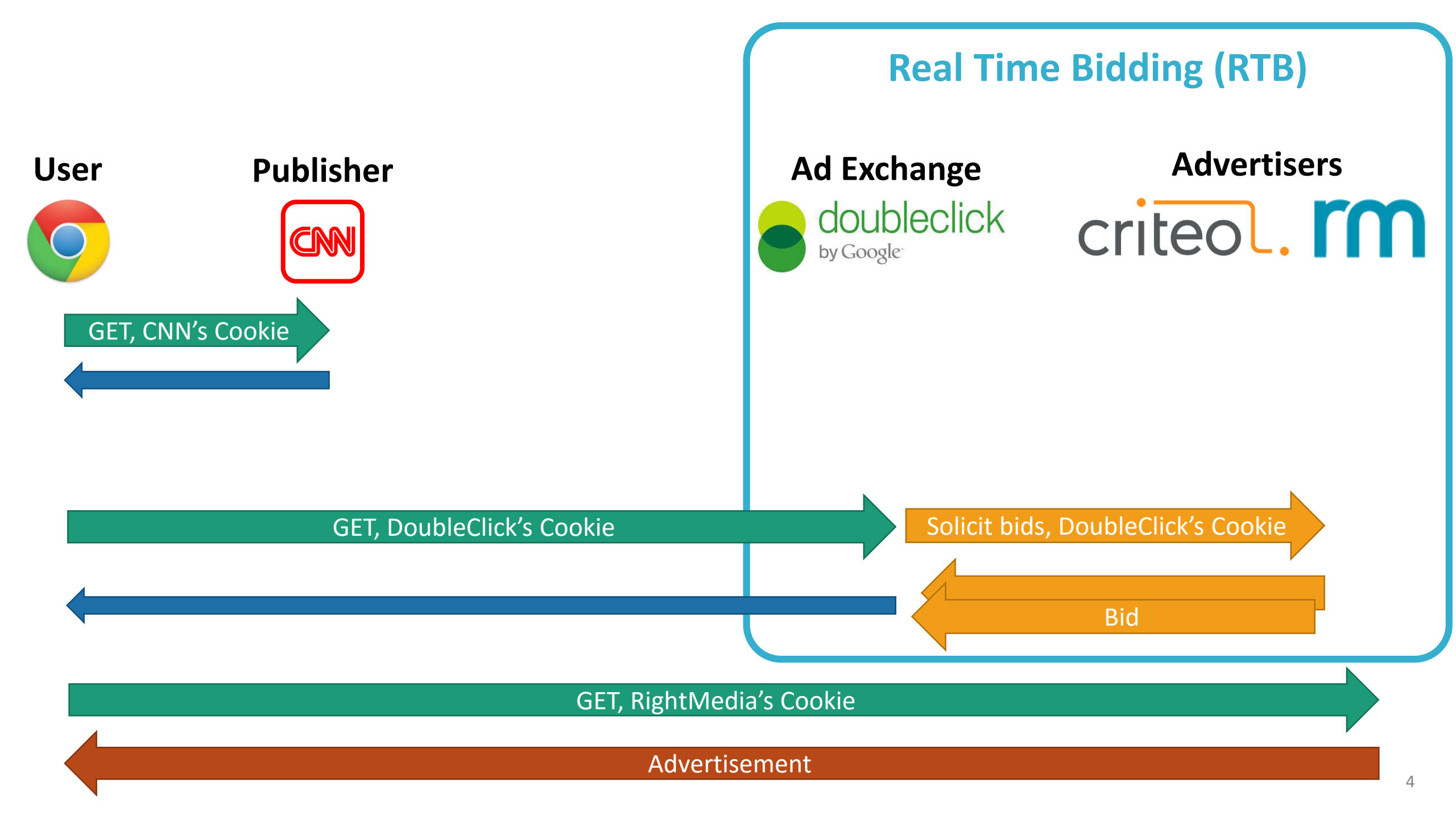
Publisher

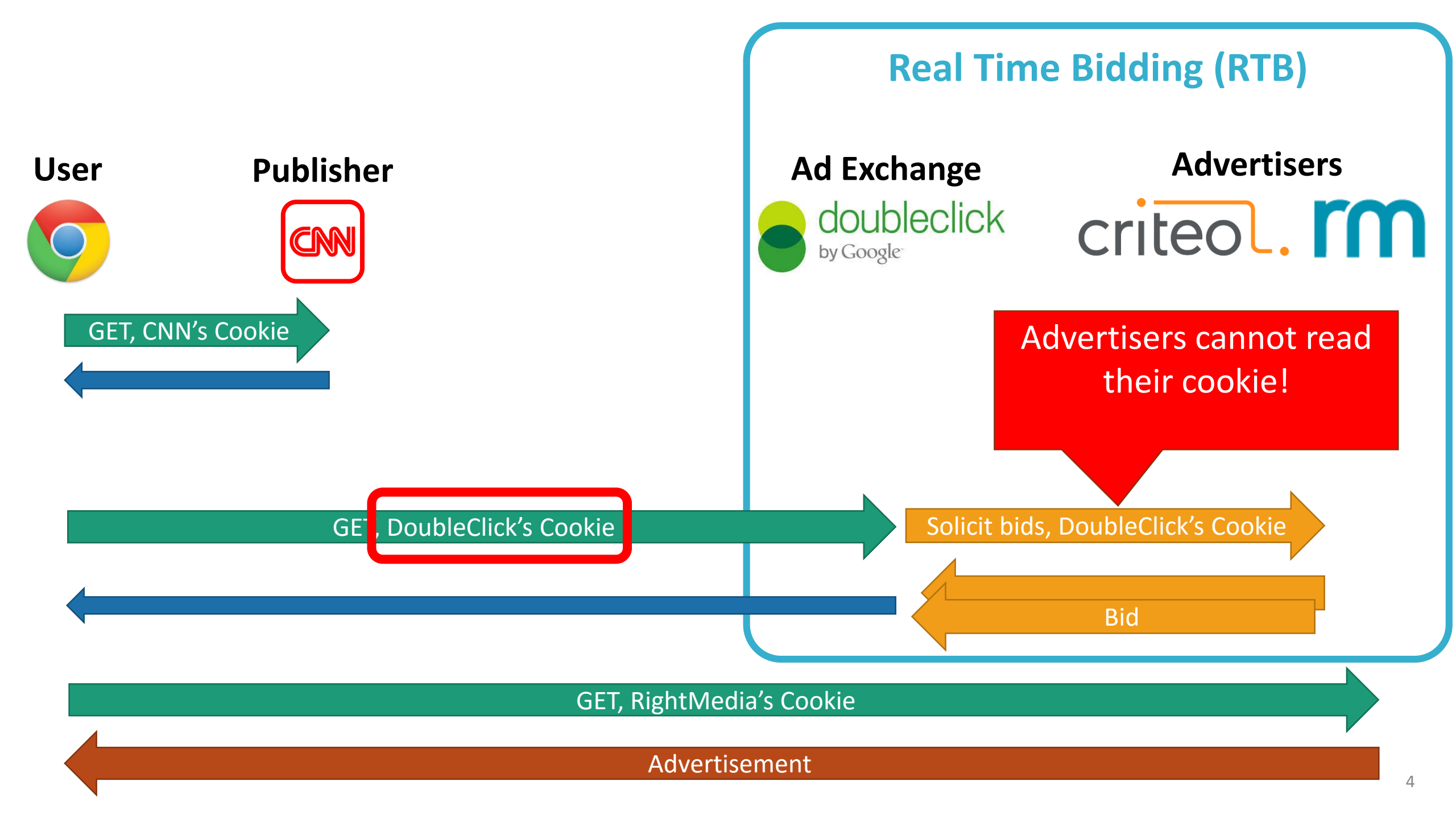
Ad Exchange

Advertisers


User
Publisher
Ad Exchange
Advertisers
doubleclick
by Google
criteo
rm
GET, CNN's Cookie
GET, DoubleClick's Cookie
Solicit bids, DoubleClick's Cookie
Bid


Real Time Bidding (RTB)
User
Publisher
Ad Exchange
Advertisers
doubleclick
by Google
criteo
rm
GET, CNN's Cookie
GET, DoubleClick's Cookie
Solicit bids, DoubleClick's Cookie
Bid


Real Time Bidding (RTB)
User
Publisher
Ad Exchange
Advertisers
doubleclick
by Google
criteo
rm
GET, CNN's Cookie
GET, DoubleClick's Cookie
Solicit bids, DoubleClick's Cookie
Bid
GET, RightMedia's Cookie
Advertisement


Real Time Bidding (RTB)
User
Publisher
Ad Exchange
Advertisers
doubleclick
by Google
criteo
rm
GET, CNN's Cookie
Advertisers cannot read their cookie!
GET, DoubleClick's Cookie
Solicit bids, DoubleClick's Cookie
Bid
GET, RightMedia's Cookie
Advertisement

Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges

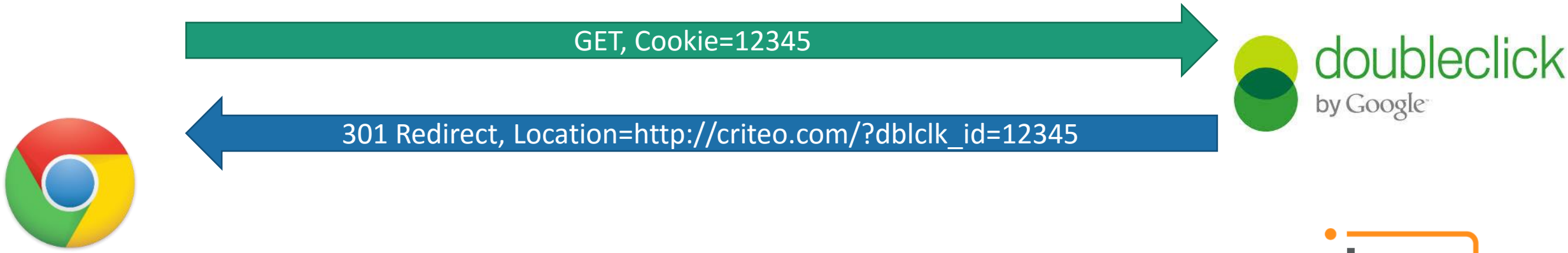
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



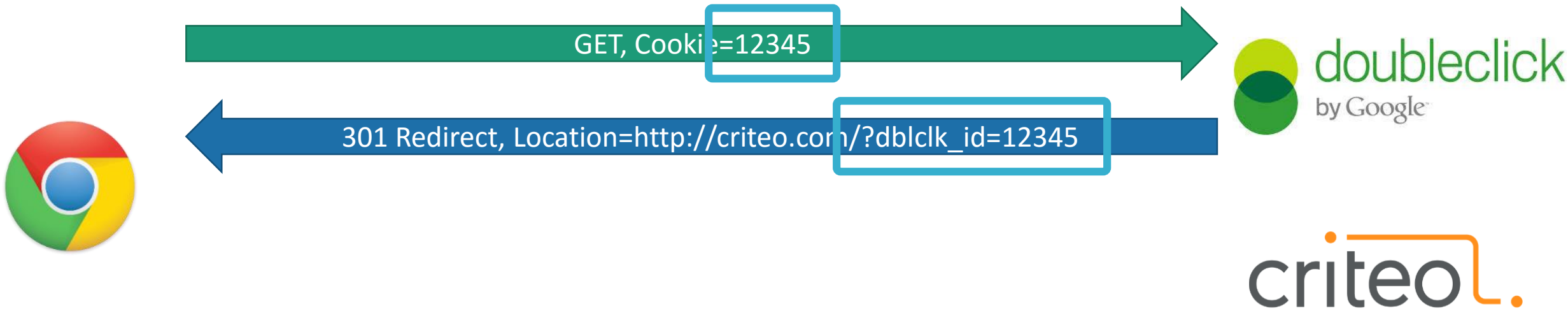
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



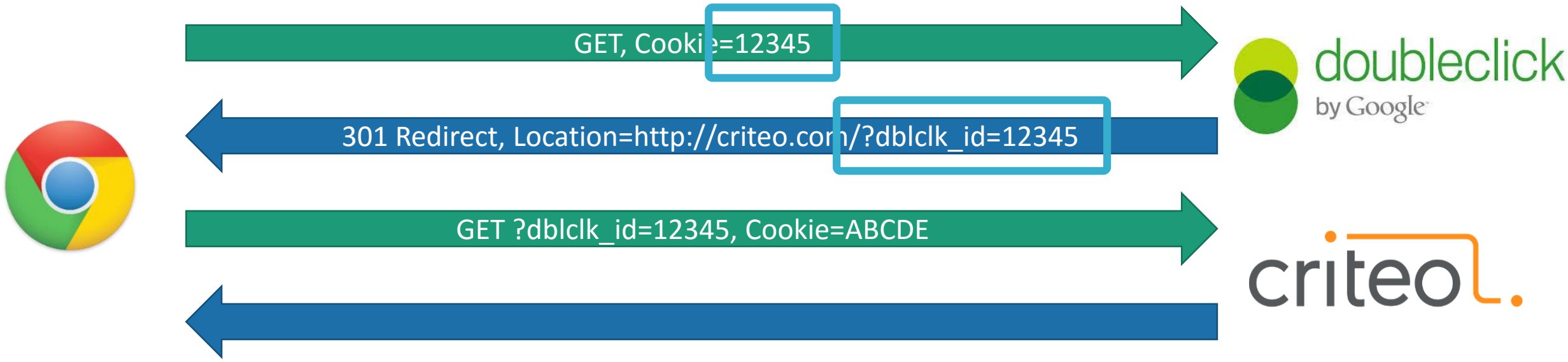
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



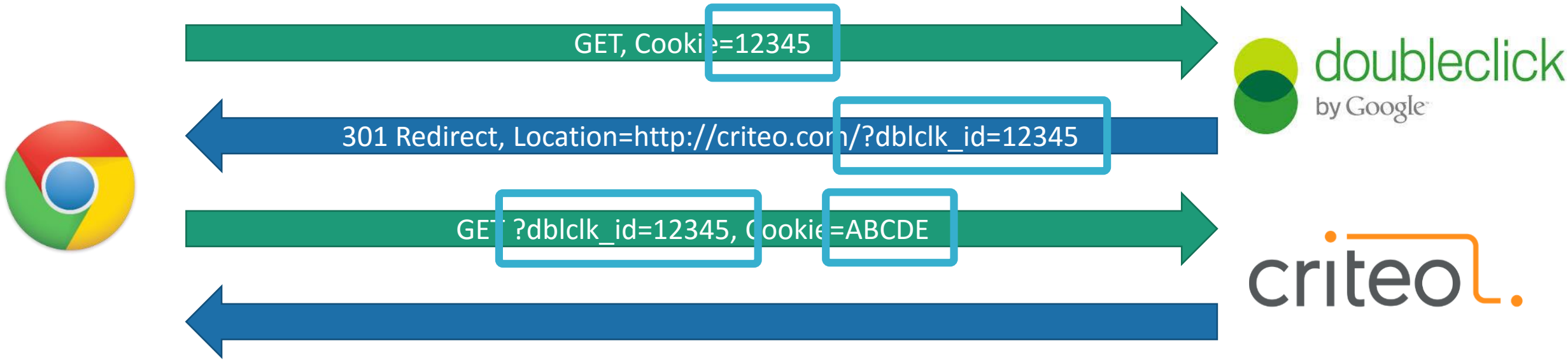
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



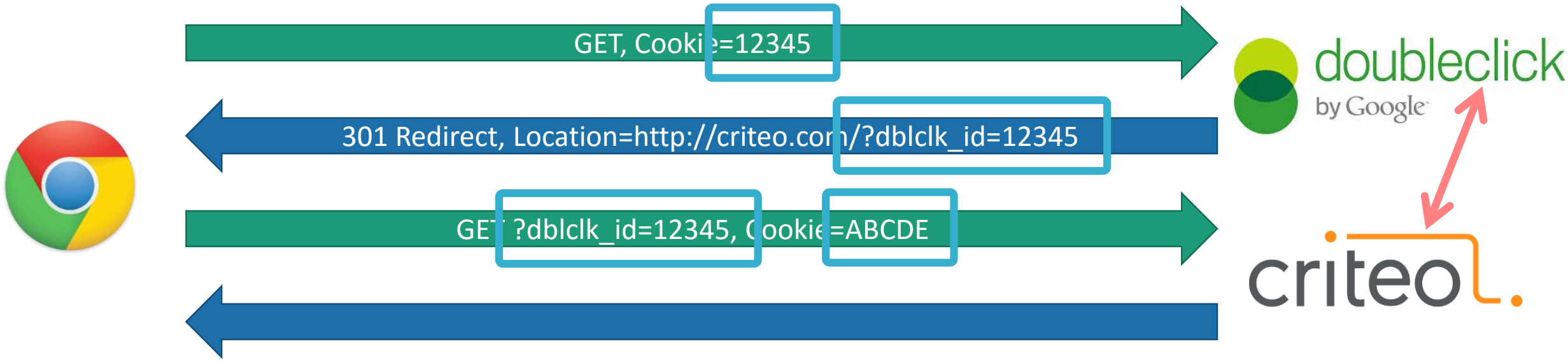
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: cookie matching

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



Prior Work

- Several studies have examined cookie matching
  - Acar *et al.* found hundreds of domains passing identifiers to each other
  - Olejnik *et al.* found 125 exchanges matching cookies
  - Falahrastegar *et al.* analyzed clusters of exchanges that share the exact same cookies

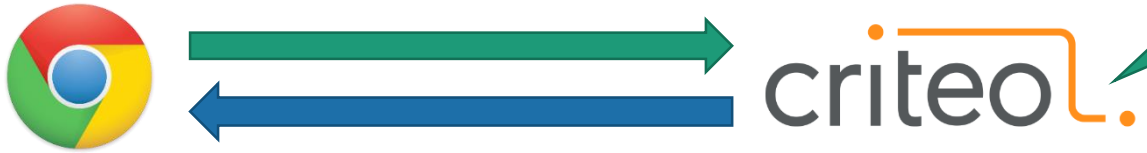
Prior Work

- Several studies have examined cookie matching
  - Acar *et al.* found hundreds of domains passing identifiers to each other
  - Olejnik *et al.* found 125 exchanges matching cookies
  - Falahrastegar *et al.* analyzed clusters of exchanges that share the exact same cookies

- These studies rely on studying HTTP requests/responses.

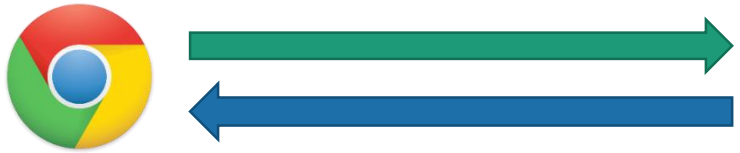
Challenge 1: Server Side Matching

Challenge 1: Server Side Matching

1) 

Criteo observes the user.
(IP: 207.91.160.7)

Challenge 1: Server Side Matching

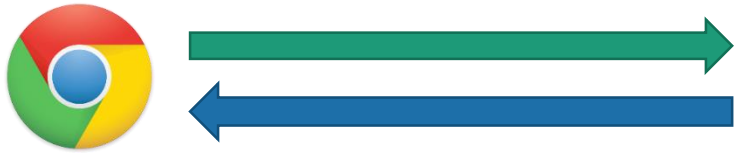
1) 

Criteo observes the user.
(IP: 207.91.160.7)

2) 

RightMedia observes the user.
(IP: 207.91.160.7)

Challenge 1: Server Side Matching

1) 

Criteo observes the user.
(IP: 207.91.160.7)

2) 

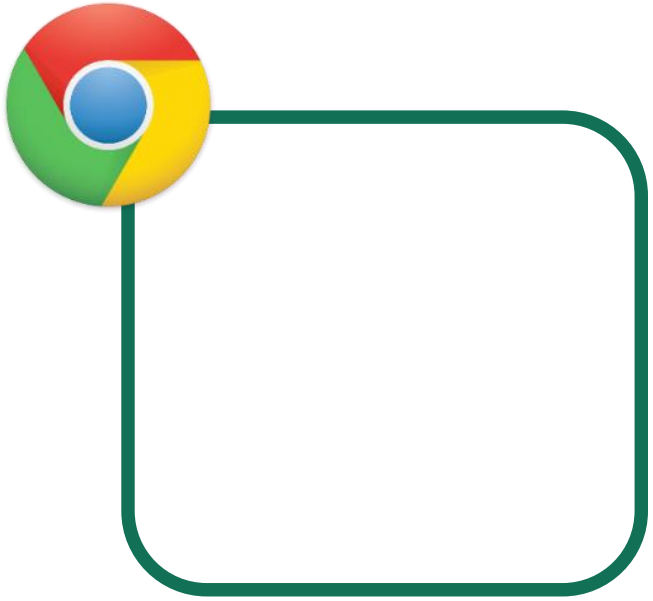
RightMedia observes the user.
(IP: 207.91.160.7)

Behind the scene, RightMedia and Criteo sync up.

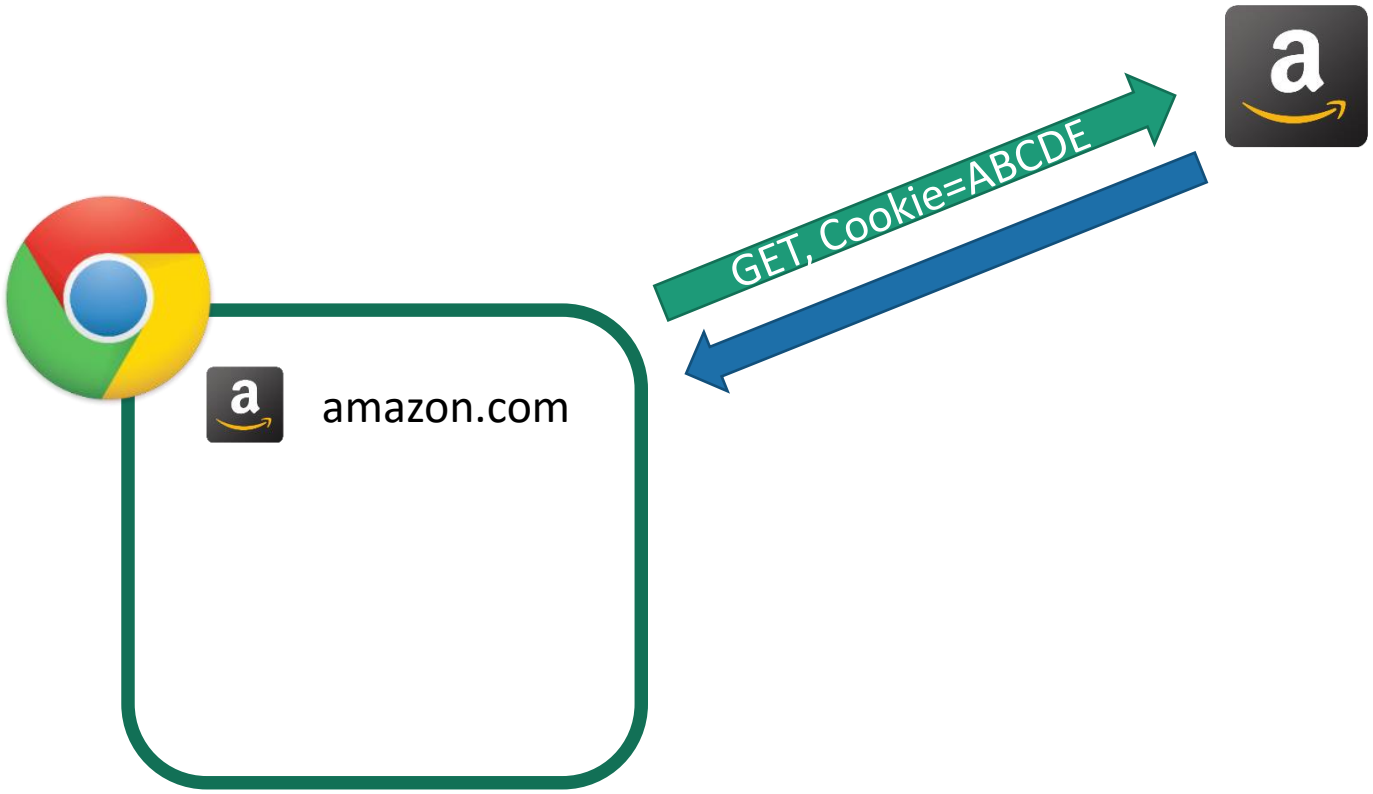
(IP: 207.91.160.7)

Challenge 2: Obfuscation

Challenge 2: Obfuscation



Challenge 2: Obfuscation



Challenge 2: Obfuscation

Challenge 2: Obfuscation

Challenge 2: Obfuscation

Challenge 2: Obfuscation

Challenge 2: Obfuscation

Challenge 2: Obfuscation

Goal

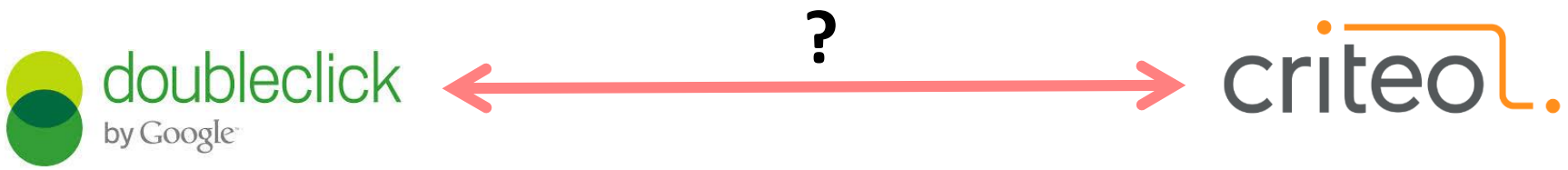
Develop a method to identify information flows (cookie matching) between ad exchanges

- Mechanism agnostic: resilient to obfuscation
- Platform agnostic: detect sharing on the client- and server-side

Goal

Develop a method to identify information flows (cookie matching) between ad exchanges

- Mechanism agnostic: resilient to obfuscation
- Platform agnostic: detect sharing on the client- and server-side



Key Insight: Use Retargeted Ads

Retargeted ads are the most highly targeted form of online ads

Cisco-Linksys AE1000 High-Performance Wireless-N Adapter

by Linksys

207 customer reviews | 10 answered questions

Price: $15.99 Prime

Only 1 left in stock.

Want it Tuesday, June 14? Order within **33 hrs 50 mins** and choose **One-Day Shipping** at checkout. Details

Sold by Home Sweet Home Direct and Fulfilled by Amazon.

Eligible for amazonsmile donation.

Want to hire a computer technician?
Buy professional computer technician services directly on Amazon. Backed by our Happiness Guarantee.
Learn more

- Networking Equipment Features: WEP Security, WPA Security, Easy Setup, WPA2

Key Insight: Use Retargeted Ads

Retargeted ads are the most highly targeted form of online ads

Cisco-Linksys AE1000 High-Performance Wireless-N Adapter

by Linksys

207 customer reviews | 10 answered questions

Price: $15.99 Prime

Only 1 left in stock.

Want it Tuesday, June 14? Order within **33 hrs 50 mins** and choose **One-Day Shipping** at checkout. Details

Sold by Home Sweet Home Direct and Fulfilled by Amazon.

Eligible for amazonsmile donation.

Want to hire a computer technician?
Buy professional computer technician services directly on Amazon. Backed by our Happiness Guarantee.
Learn more

- Networking Equipment Features: WEP Security, WPA Security, Easy Setup, WPA2

Key Insight: Use Retargeted Ads

Retargeted ads are the most highly targeted form of online ads

Cisco-Linksys AE1000 High-Performance Wireless-N Adapter

by Linksys

207 customer reviews | 10 answered questions

Price: $15.99 Prime

Only 1 left in stock.

Want it Tuesday, June 14? Order within **33 hrs 50 mins** and choose **One-Day Shipping** at checkout. Details

Sold by Home Sweet Home Direct and Fulfilled by Amazon.

Eligible for amazonsmile donation.

Want to hire a computer technician?
Buy professional computer technician services directly on Amazon. Backed by our Happiness Guarantee.
Learn more

- Networking Equipment Features: WEP Security, WPA Security, Easy Setup, WPA2

Key insight: because retargets are so specific, they can be used to conduct controlled experiments

- Information **must be** shared between ad exchanges to serve retargeted ads

Contributions

1. Novel methodology for identifying information flows between ad exchanges

2. Demonstrate the impact of ad network obfuscation in practice
   - 31% of cookie matching partners cannot be identified using heuristics

3. Develop a method to categorize information sharing relationships

4. Use graph analysis to infer the roles of actors in the ad ecosystem

Contributions

1. Novel methodology for identifying information flows between ad exchanges

2. Demonstrate the impact of ad network obfuscation in practice
   - 31% of cookie matching partners cannot be identified using heuristics

3. Develop a method to categorize information sharing relationships

4. ~~Use graph analysis to infer the roles of actors in the ad ecosystem~~

Data Collection
Classifying Ad Network Flows
Results

Using Retargets as an Experimental Tool

Key observation: retargets are only served under very specific circumstances

1)   criteo

Using Retargets as an Experimental Tool

Key observation: retargets are only served under very specific circumstances

1)

Using Retargets as an Experimental Tool

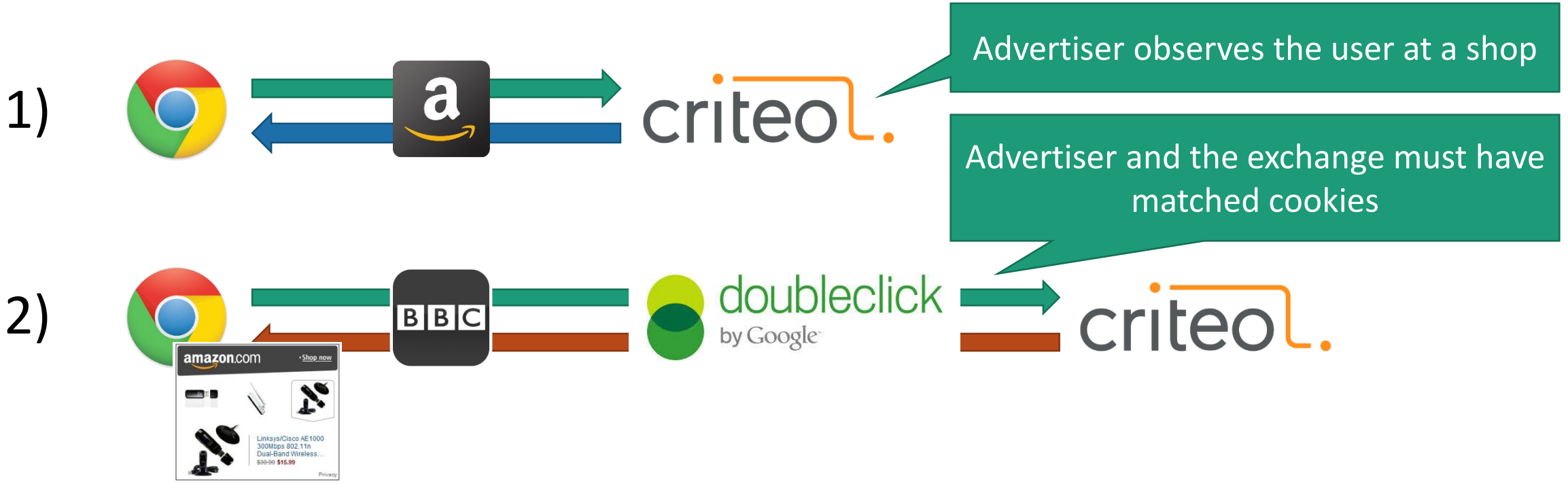
Key observation: retargets are only served under very specific circumstances

1)

2)

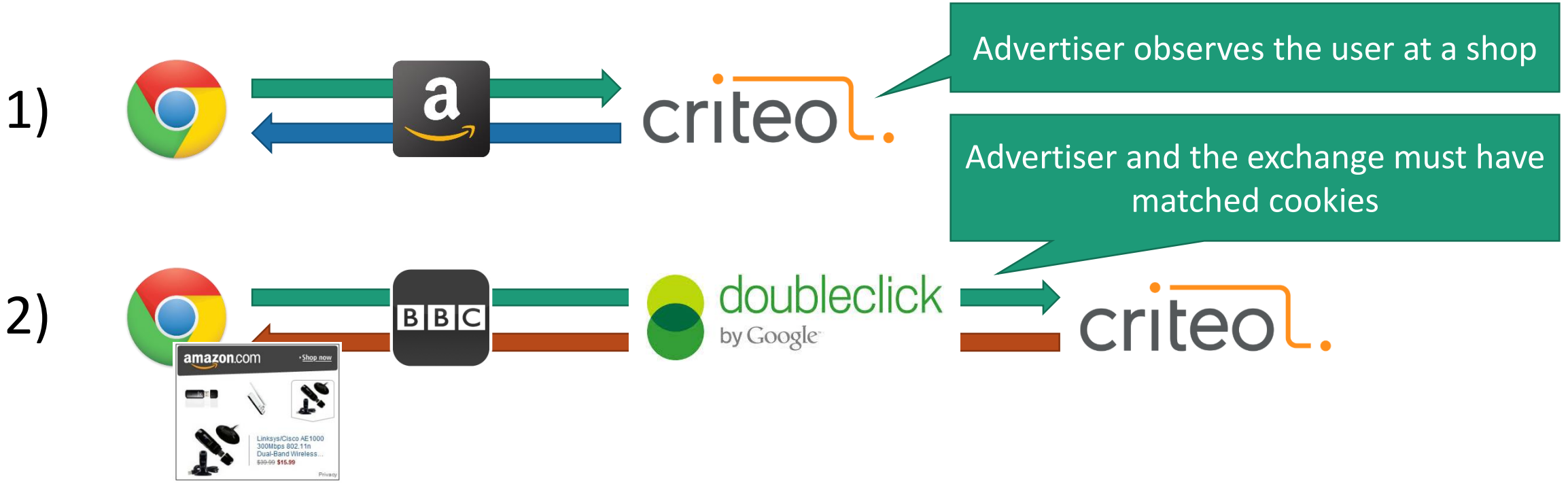
Using Retargets as an Experimental Tool

Key observation: retargets are only served under very specific circumstances



Using Retargets as an Experimental Tool

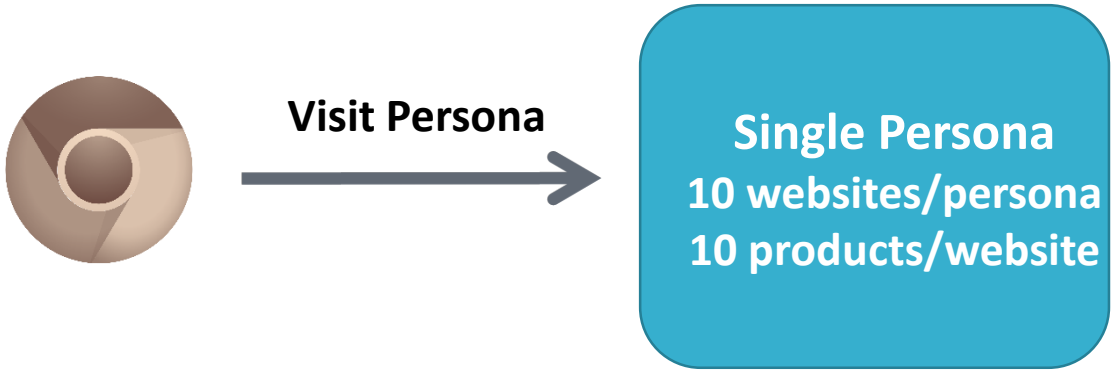
Key observation: retargets are only served under very specific circumstances



This implies a causal flow of information from Exchange → Advertiser

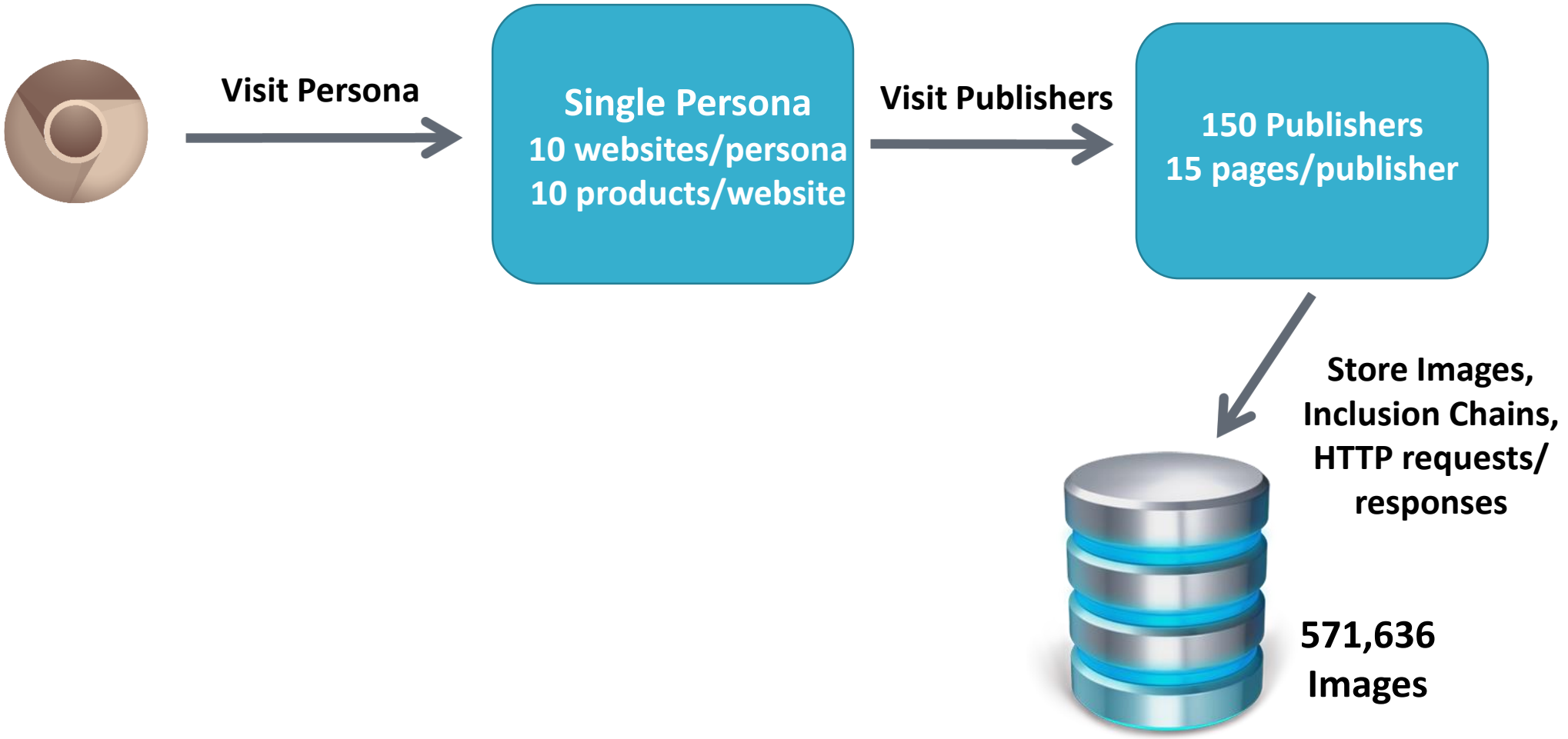
Data Collection Overview

Data Collection Overview

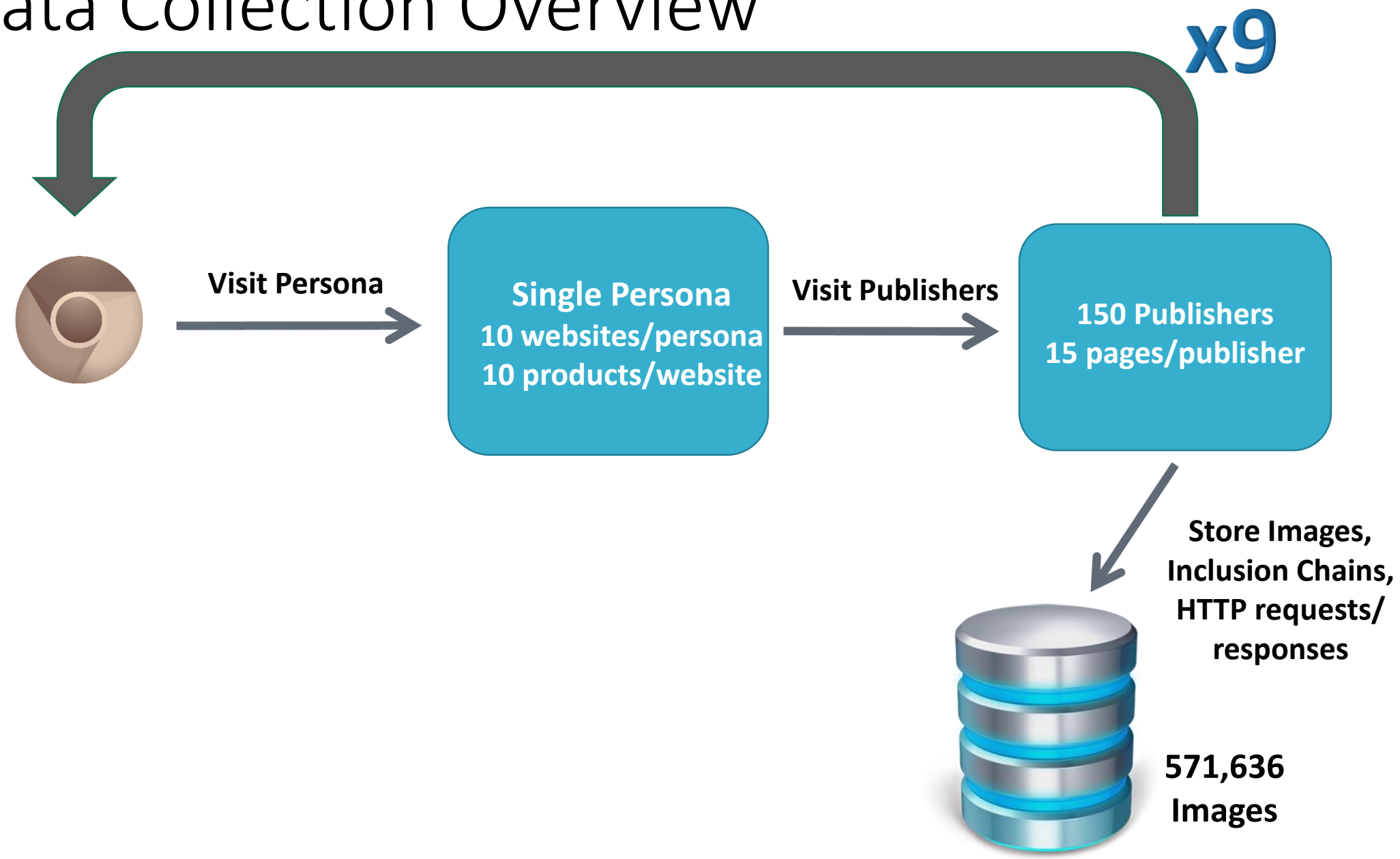


Data Collection Overview

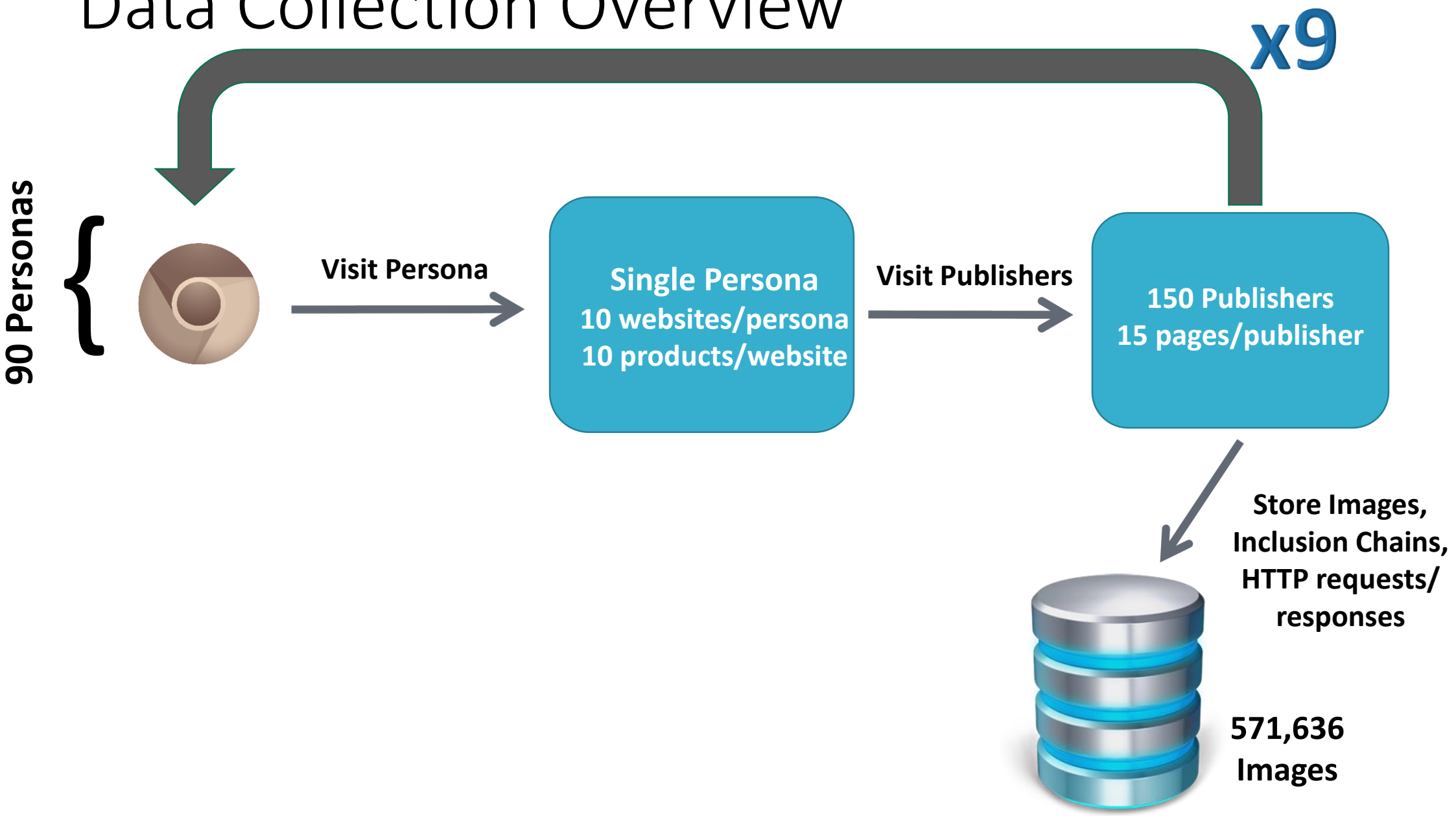
Data Collection Overview



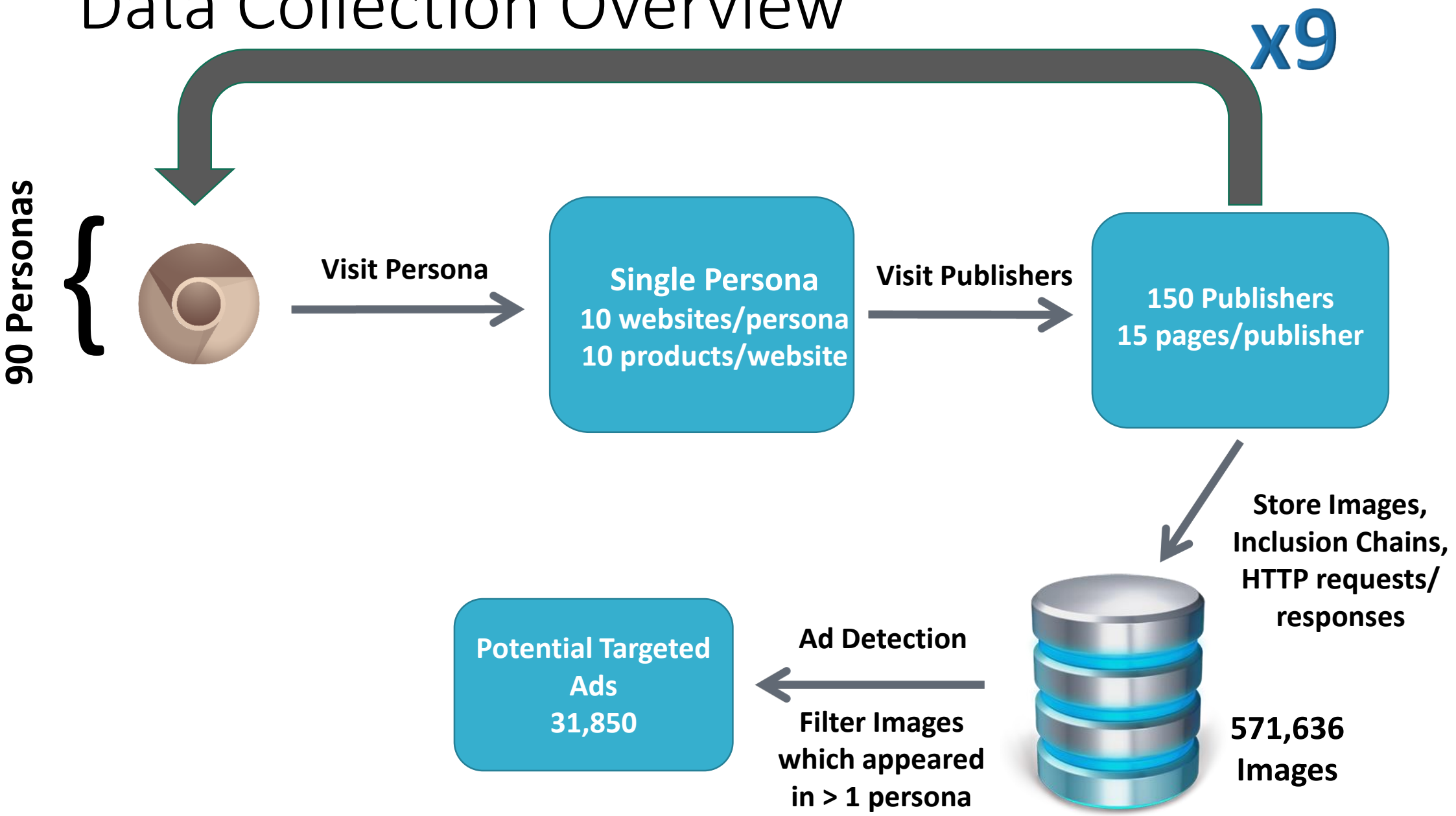
Data Collection Overview



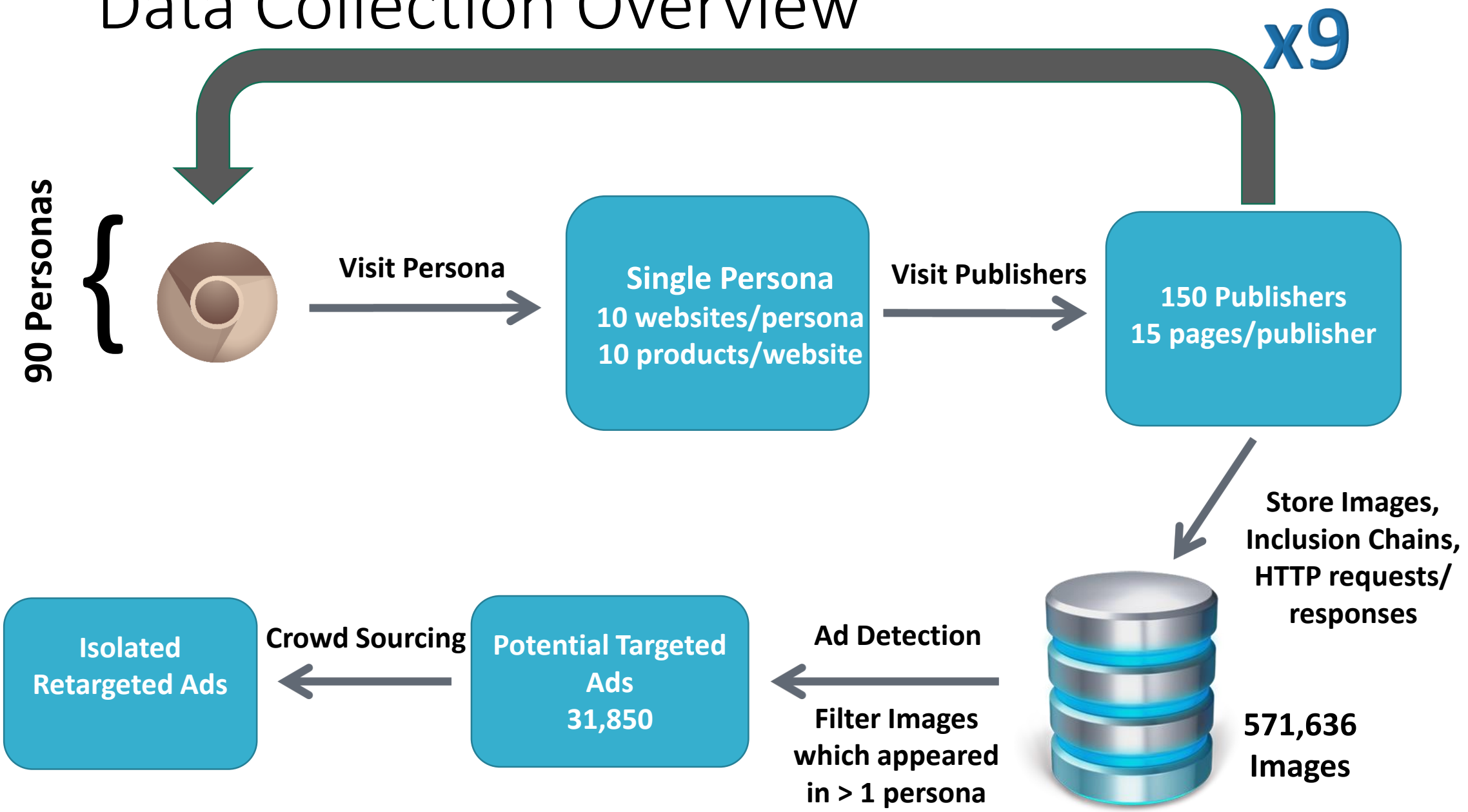
Data Collection Overview



Data Collection Overview



Data Collection Overview



Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

Crowd Sourcing

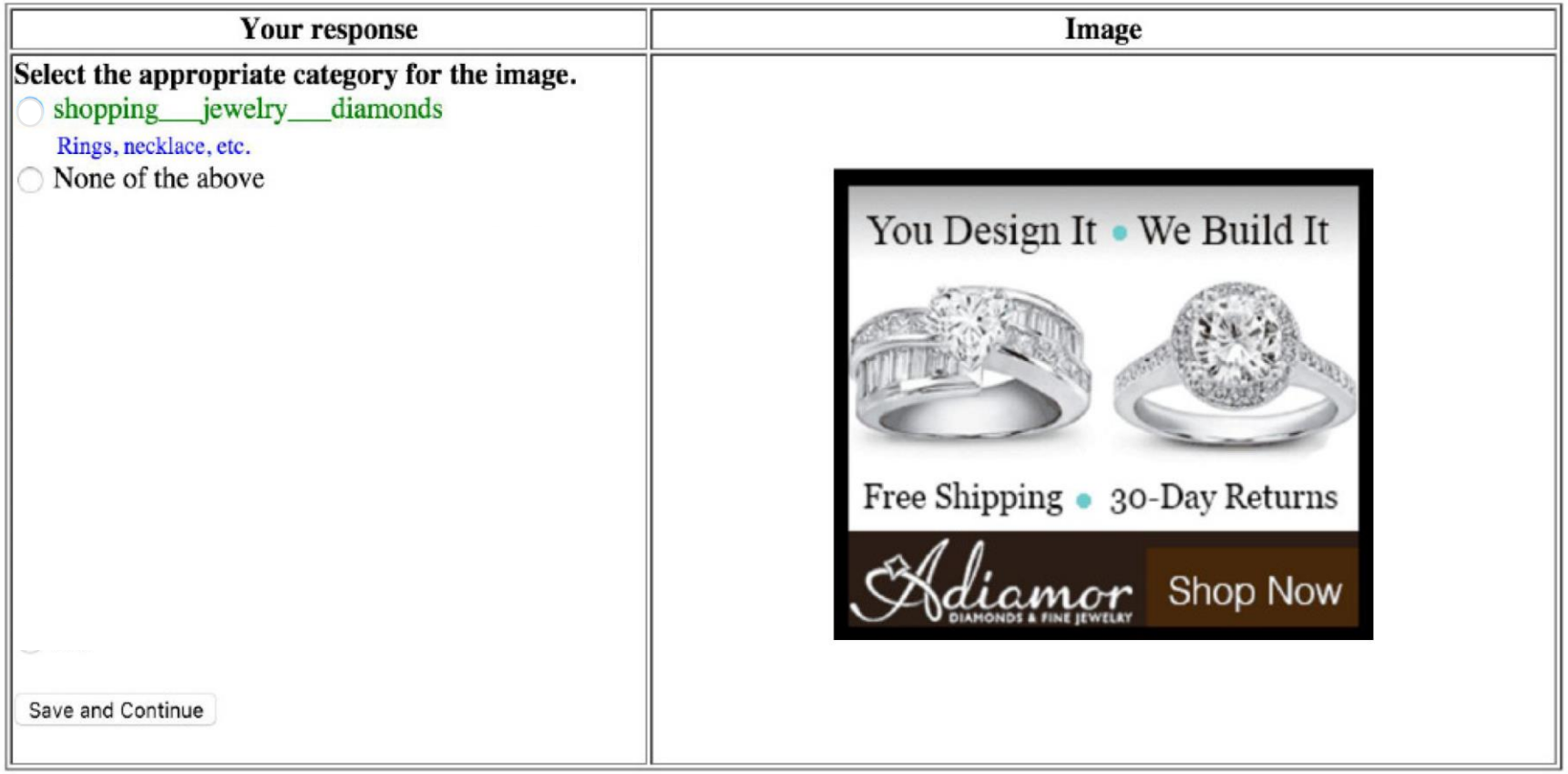
We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- $415 spent.

Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- $415 spent.



Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- $415 spent.

Your response	Image
Select the appropriate category for the image. (•) shopping___jewelry___diamonds Rings, necklace, etc. ( ) None of the above **Does this image say it came from one the following websites?** ( ) adiamor.com ( ) bluenile.com ( ) gilletts.com.au ( ) lumeradiamonds.com ( ) shaneco.com ( ) szul.com ( ) totaram.com ( ) washingtondiamond.com ( ) whiteflash.com ( ) No Save and Continue	

Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- $415 spent.

Your response	Image
Select the appropriate category for the image. (•) shopping___jewelry___diamonds Rings, necklace, etc. ( ) None of the above **Does this image say it came from one the following websites?** (•) adiamor.com ( ) bluenile.com ( ) gilletts.com.au ( ) lumeradiamonds.com ( ) shaneco.com ( ) szul.com ( ) totaram.com ( ) washingtondiamond.com ( ) whiteflash.com ( ) No Save and Continue	

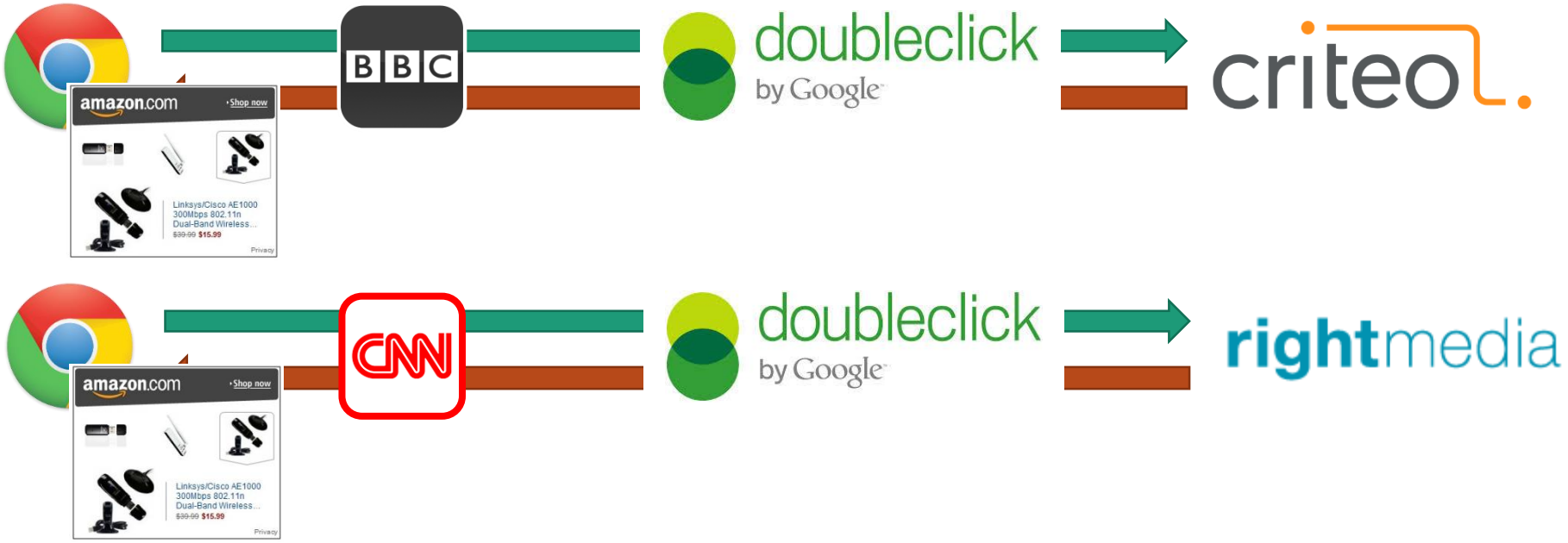
Final Dataset

5,102 unique retargeted ads

- From 281 distinct online retailers

35,448 publisher-side chains that served the retargets

- We observed some retargets multiple times



Data Collection

Classifying Ad Network Flows Results

A look at Publisher Chains

A look at Publisher Chains

Example

Publisher-side chain

A look at Publisher Chains

Example

Shopper-side chain

Publisher-side chain

A look at Publisher Chains

Example

Shopper-side chain

Publisher-side chain

- How does Criteo know to serve ad on BBC?

A look at Publisher Chains

Example

Shopper-side chain

Publisher-side chain

- How does Criteo know to serve ad on BBC?
  - In this case it is pretty trivial.
  - Criteo observed us on the shopper.

A look at Publisher Chains

Example

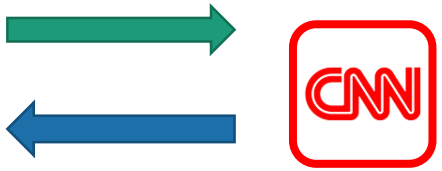
Shopper-side chain

Publisher-side chain

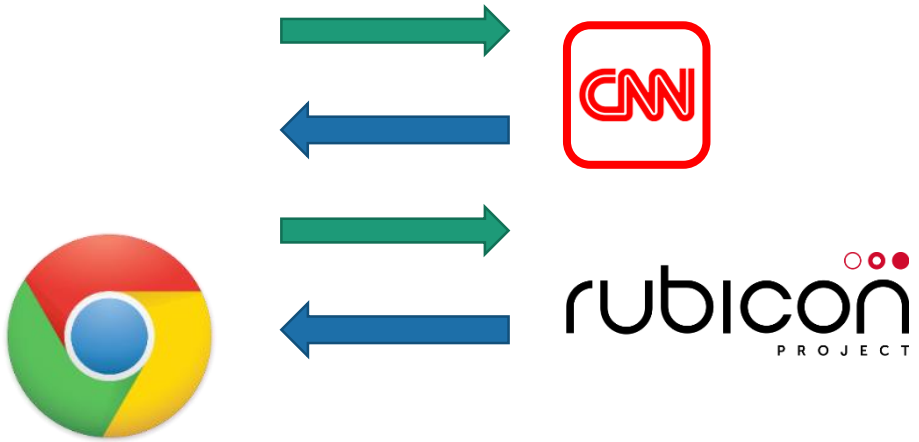
- How does Criteo know to serve ad on BBC?
  - In this case it is pretty trivial.
  - Criteo observed us on the shopper.
- Can we classify all such publisher-side chains?

What is a Chain?

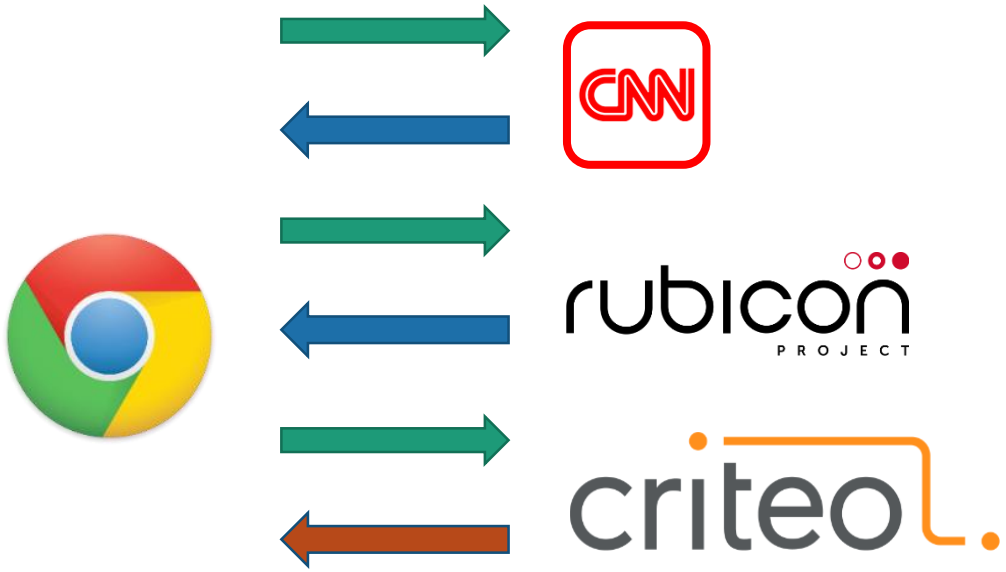
What is a Chain?



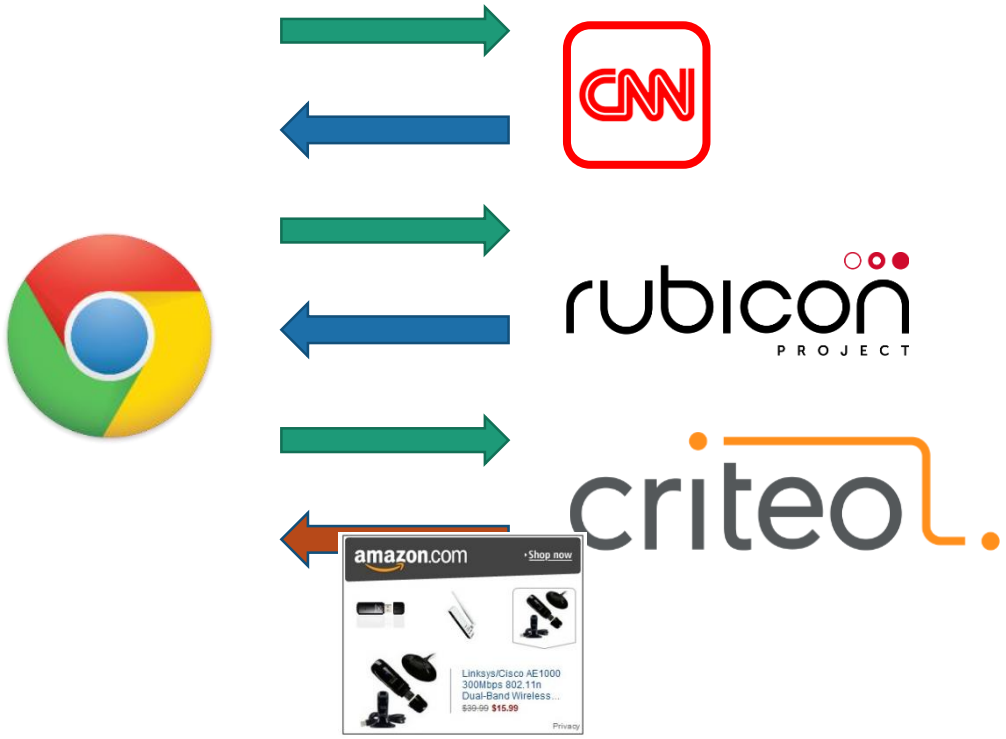
What is a Chain?



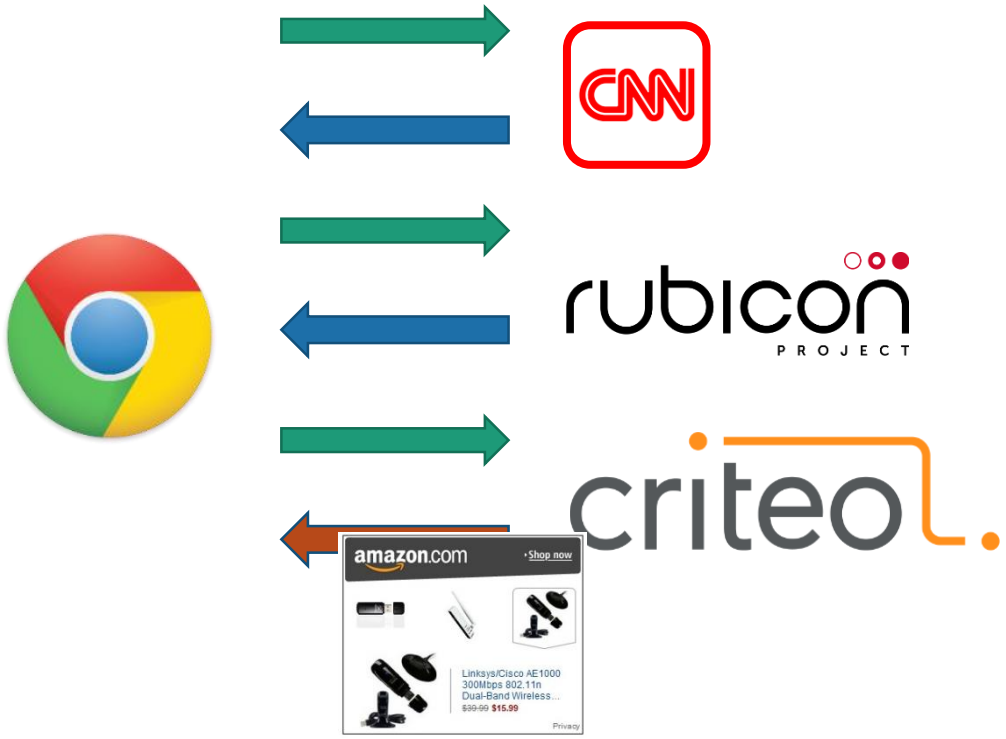
What is a Chain?



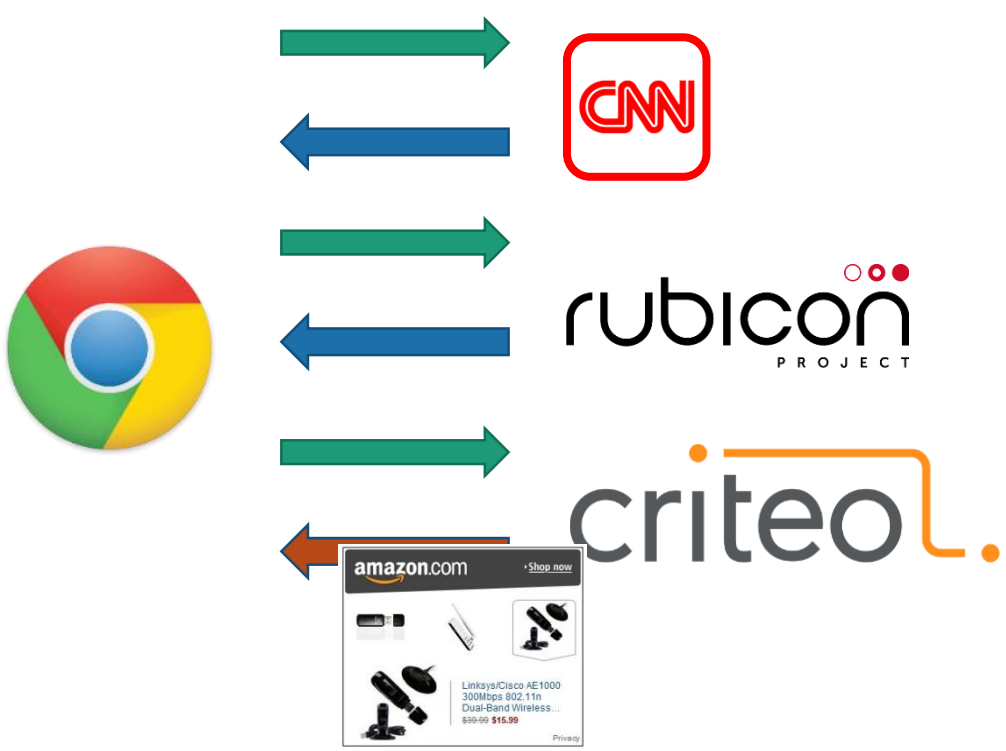
What is a Chain?

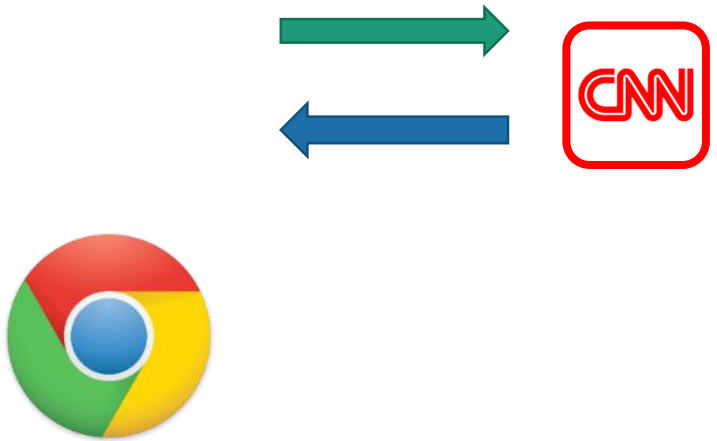


What is a Chain?

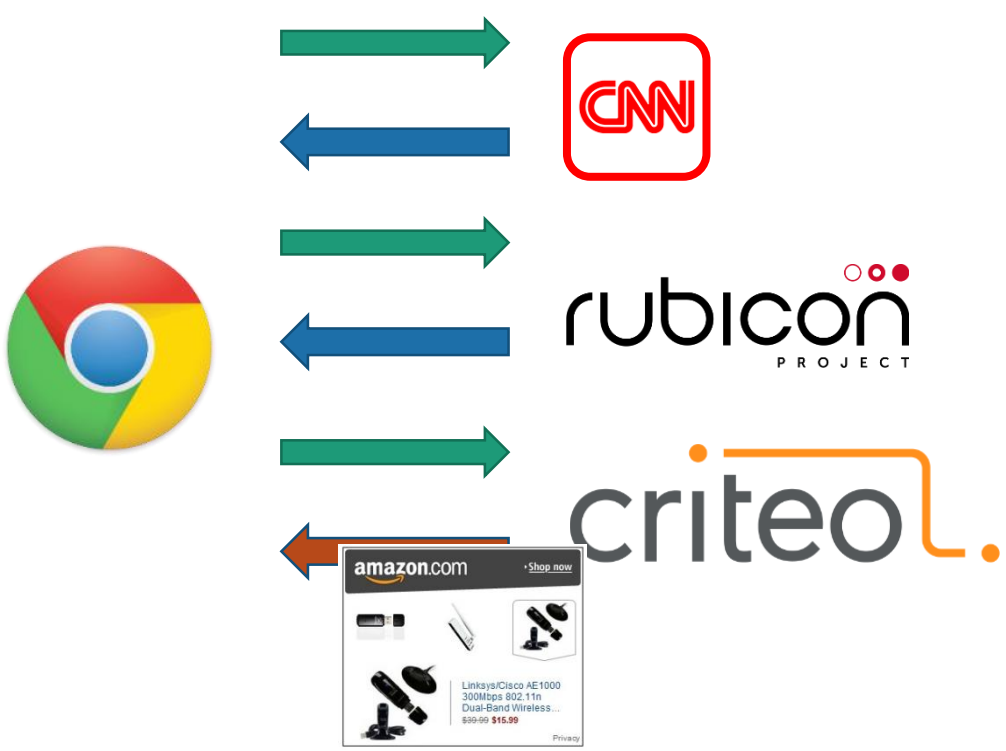


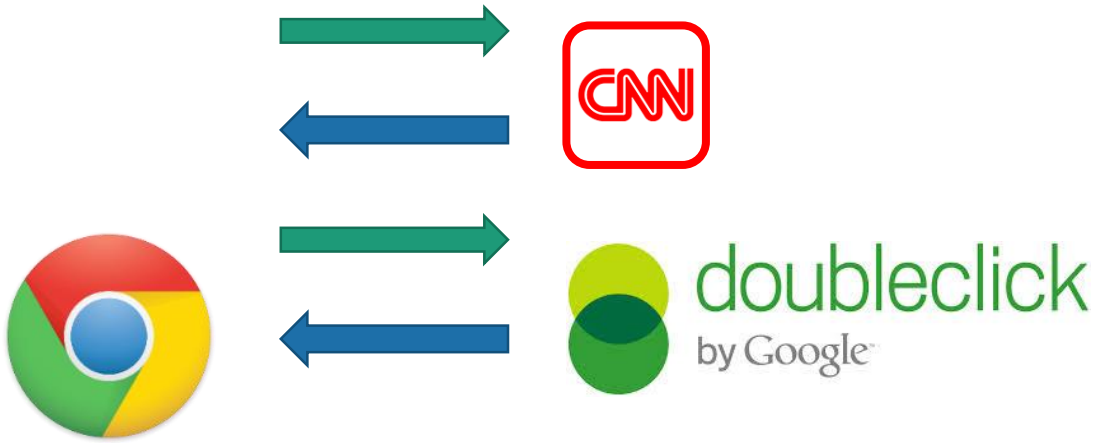
What is a Chain?



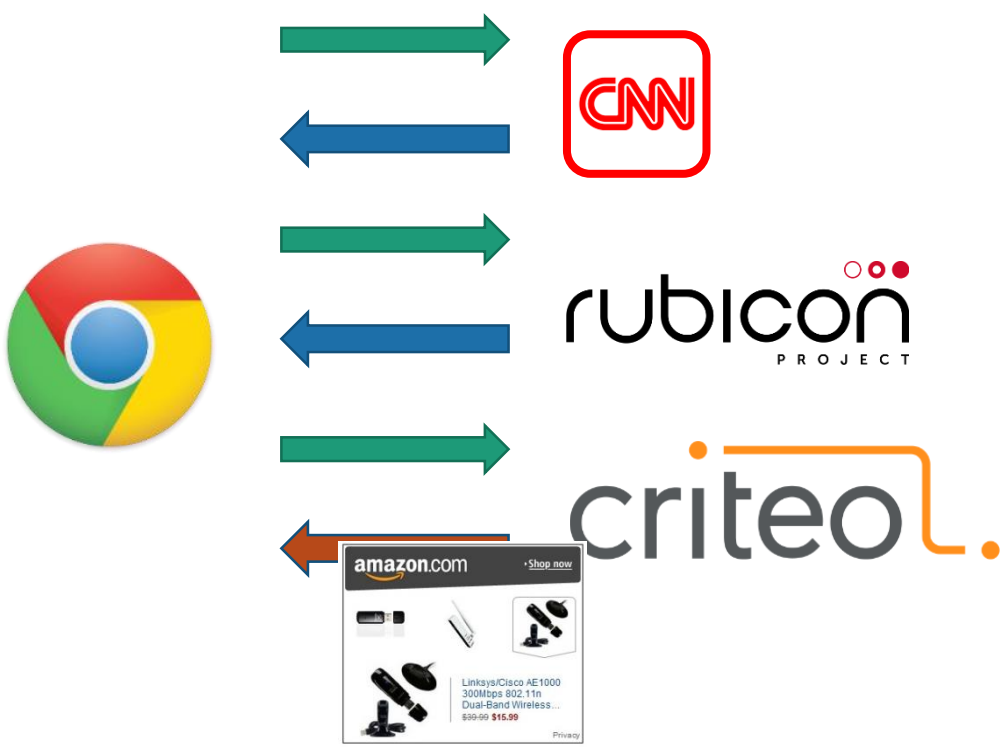


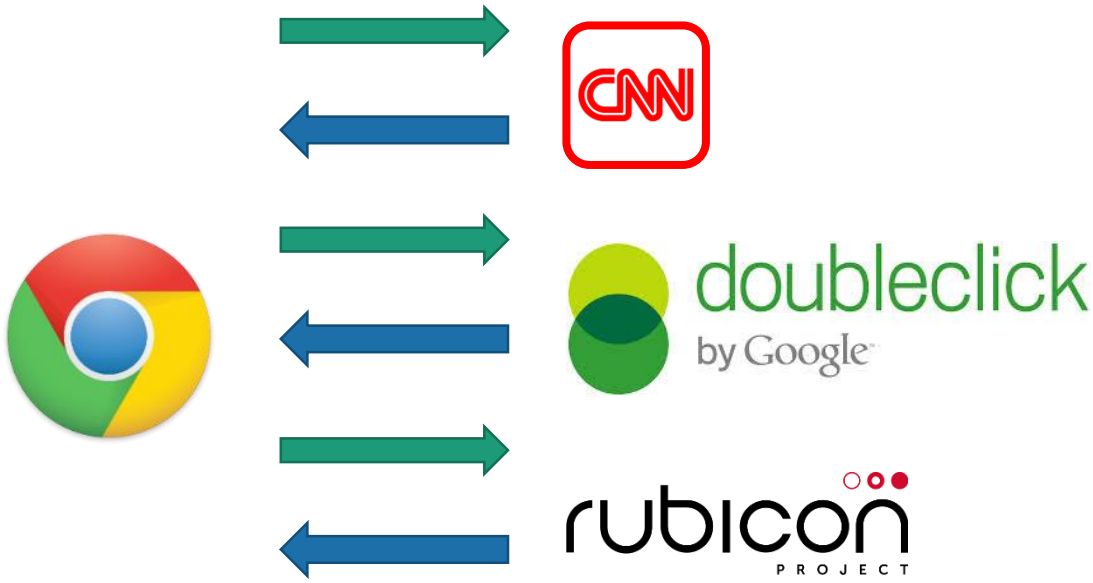
What is a Chain?



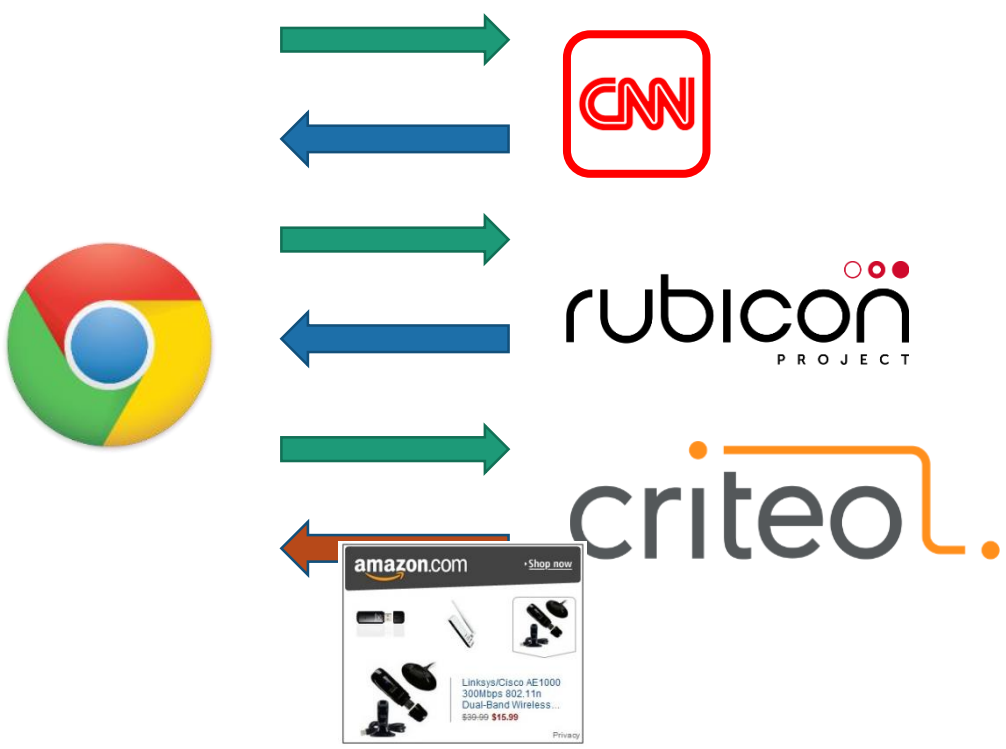


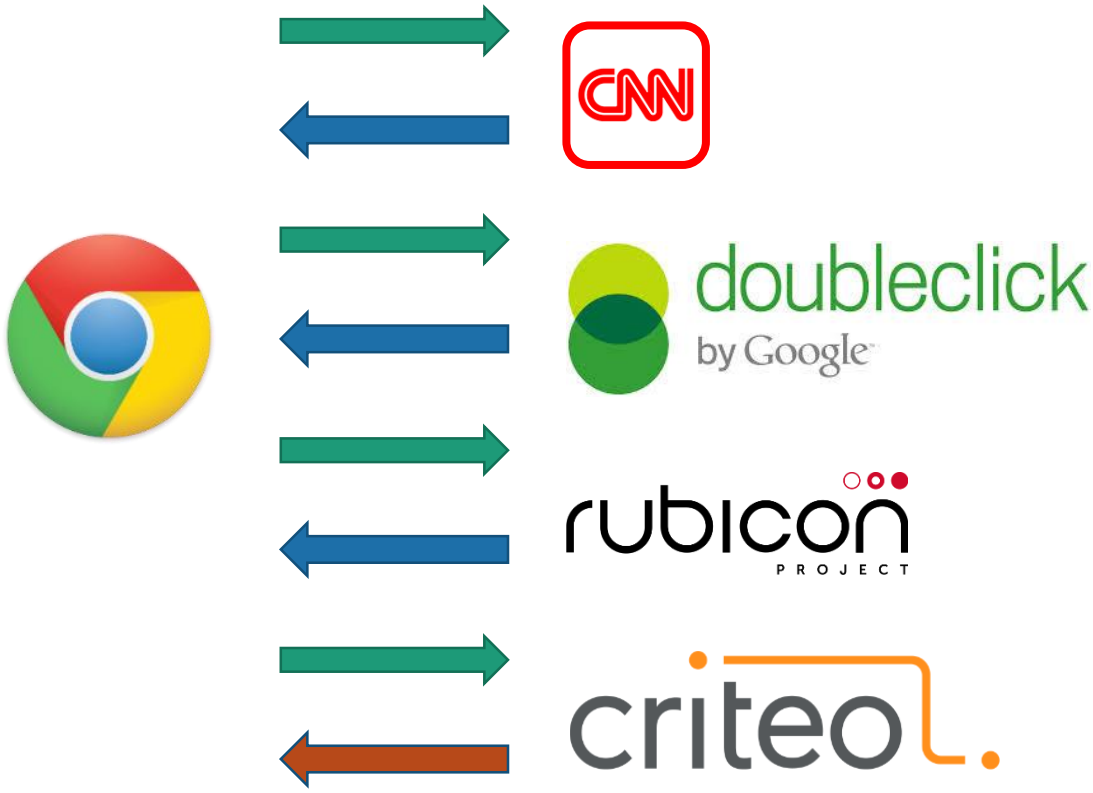
What is a Chain?



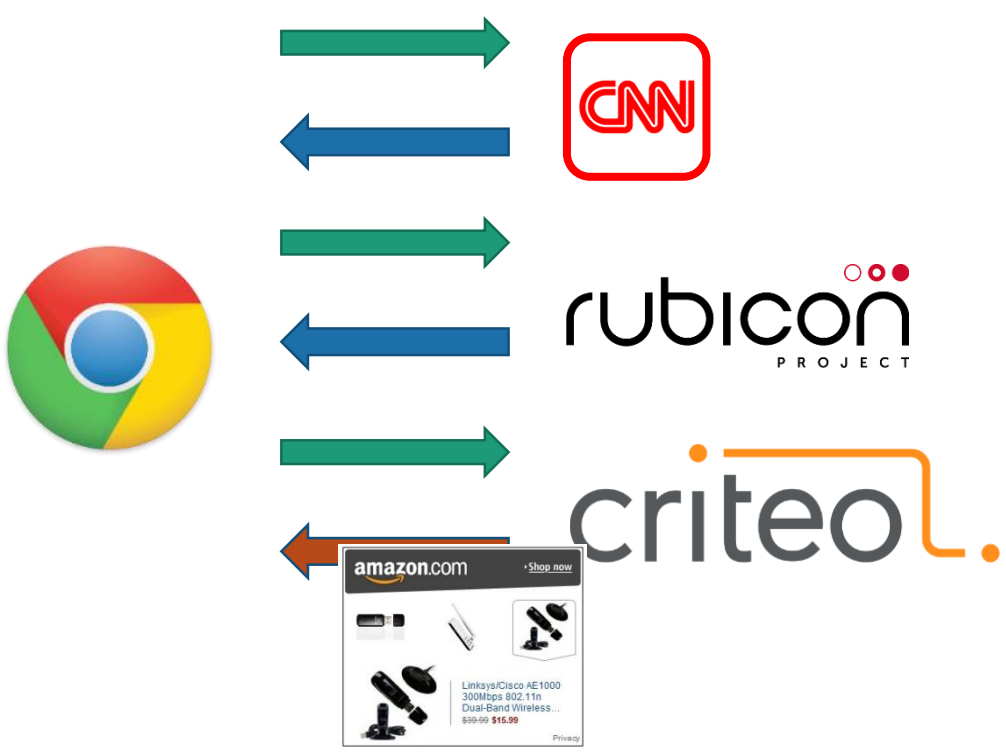


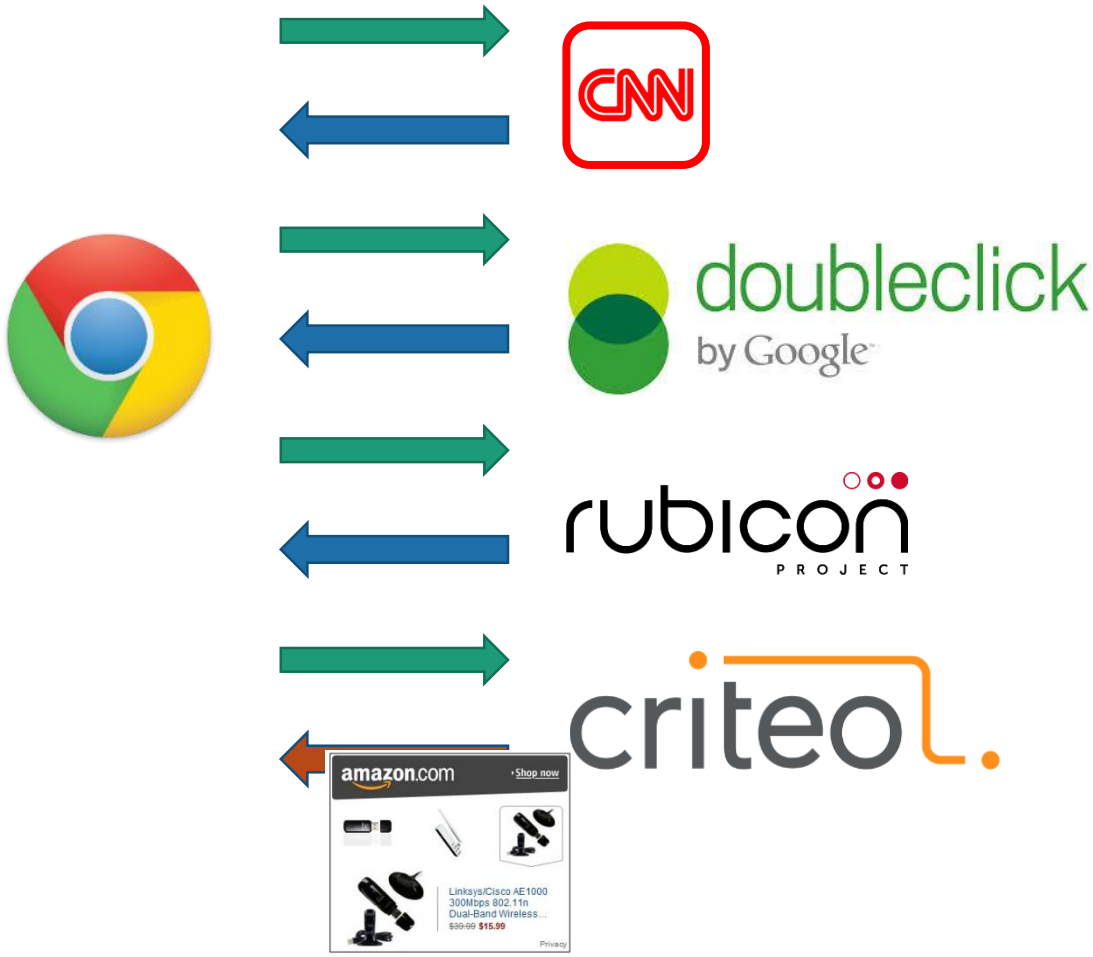
What is a Chain?



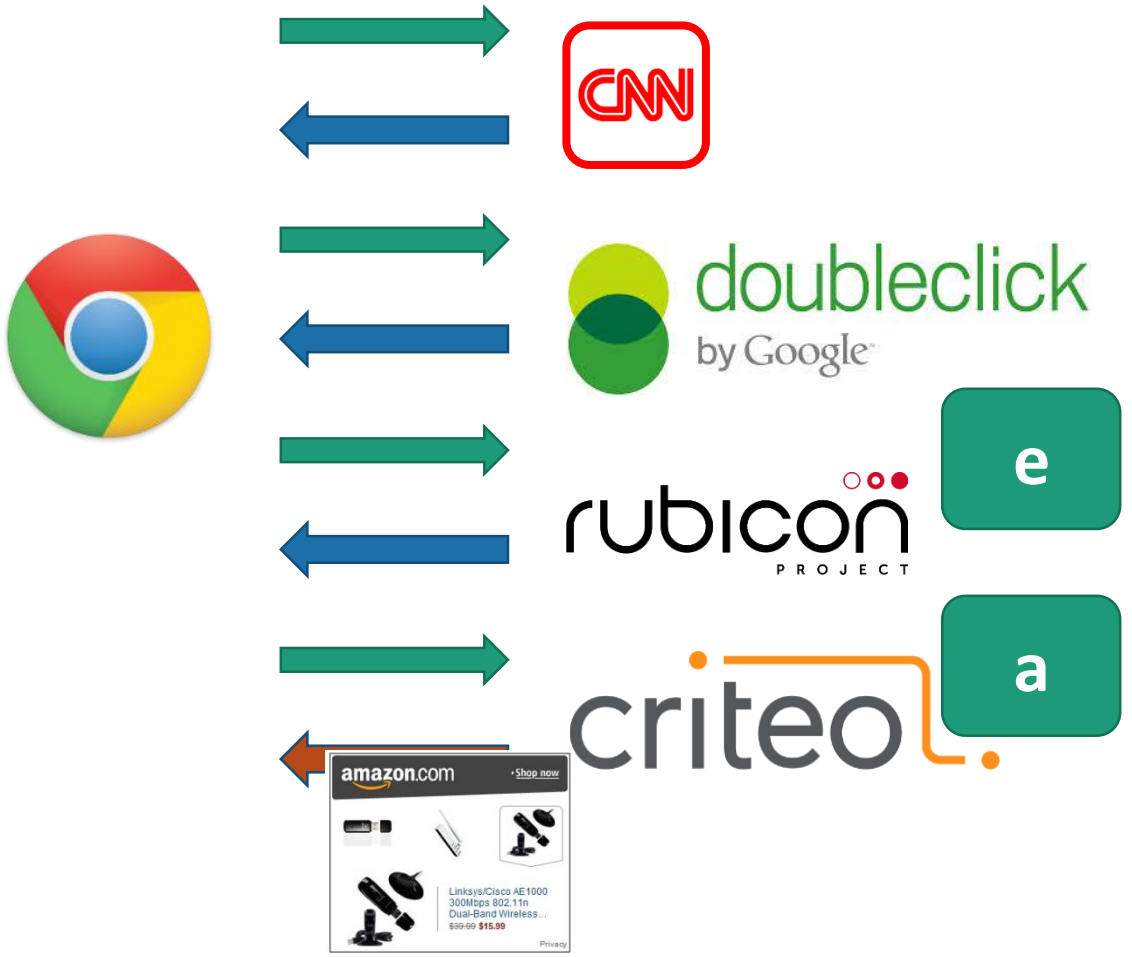


What is a Chain?

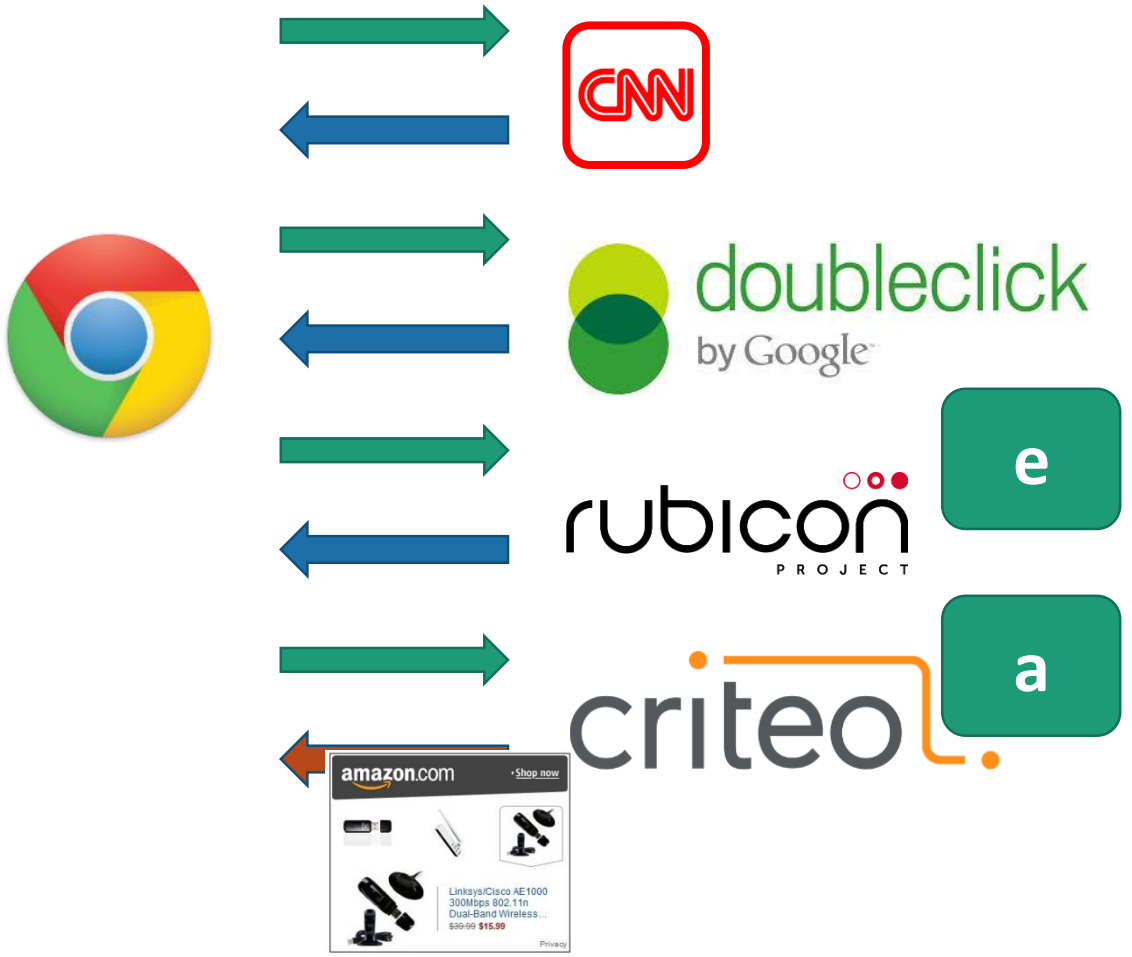


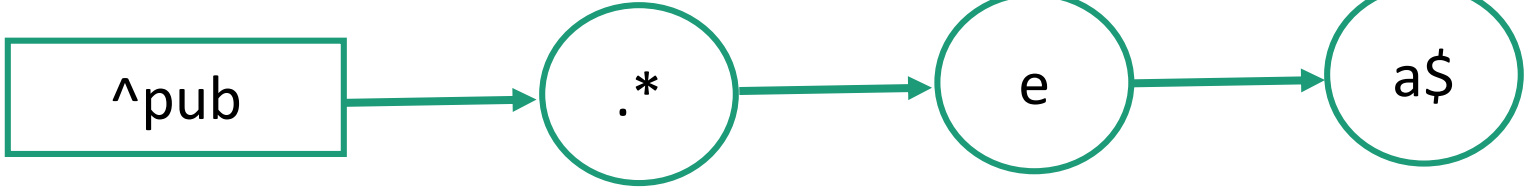


What is a Chain?



What is a Chain?





Four Classifications

Four possible ways for a retargeted ad to be served

1. Direct (Trivial) Matching
2. Cookie Matching
3. Indirect Matching
4. Latent (Server-side) Matching

Four Classifications

Four possible ways for a retargeted ad to be served

1. Direct (Trivial) Matching
2. Cookie Matching
3. ~~Indirect Matching~~
4. Latent (Server-side) Matching

1) Direct (Trivial) Matching

Shopper-side

Publisher-side

Example

Rule

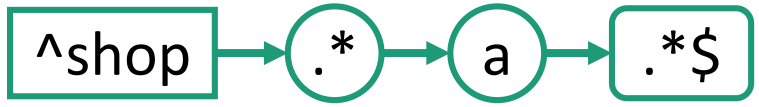
1) Direct (Trivial) Matching

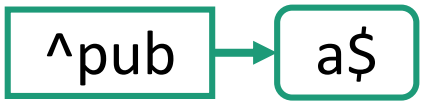
Shopper-side

Publisher-side

Example

Rule





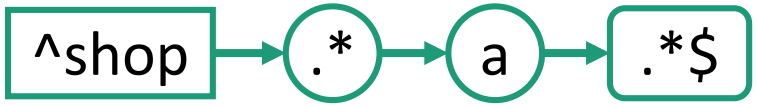
1) Direct (Trivial) Matching

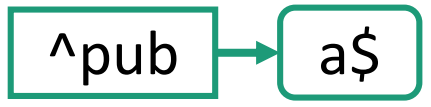
Shopper-side

Publisher-side

Example

Rule





a is the advertiser that serves the retarget

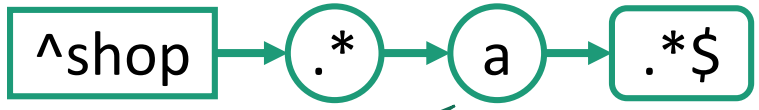
1) Direct (Trivial) Matching

Shopper-side

Publisher-side

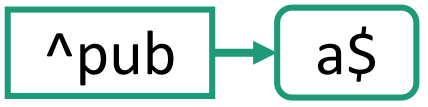
Example

Rule



a must appear on the shopper-side…

… but other trackers may also appear



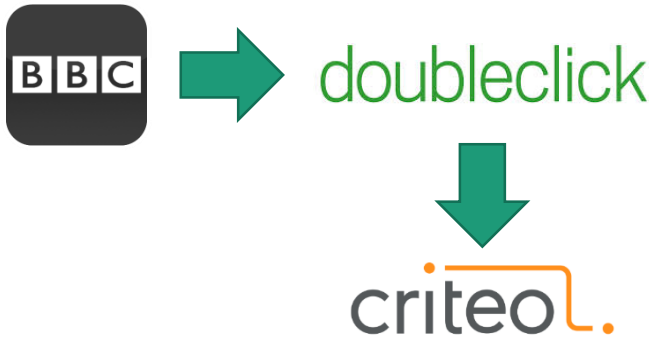
a is the advertiser that serves the retarget

2) Cookie Matching

Shopper-side

Publisher-side

Example



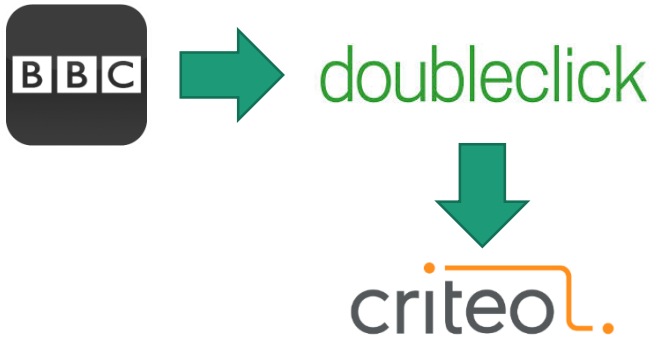
Rule

2) Cookie Matching

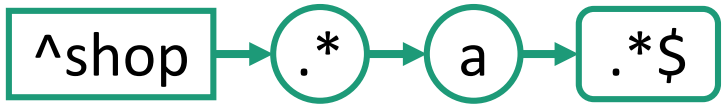
Shopper-side

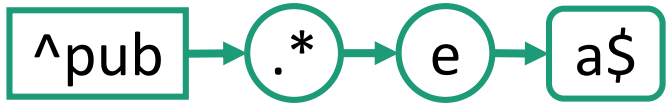
Publisher-side

Example



Rule



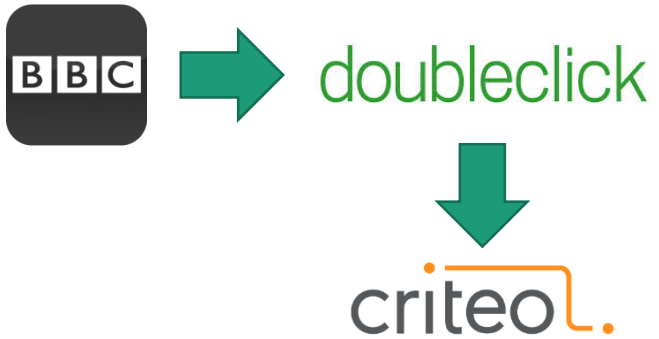


2) Cookie Matching

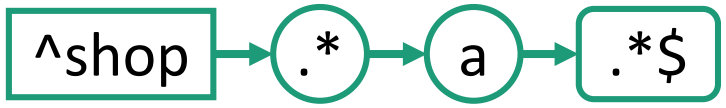
Shopper-side

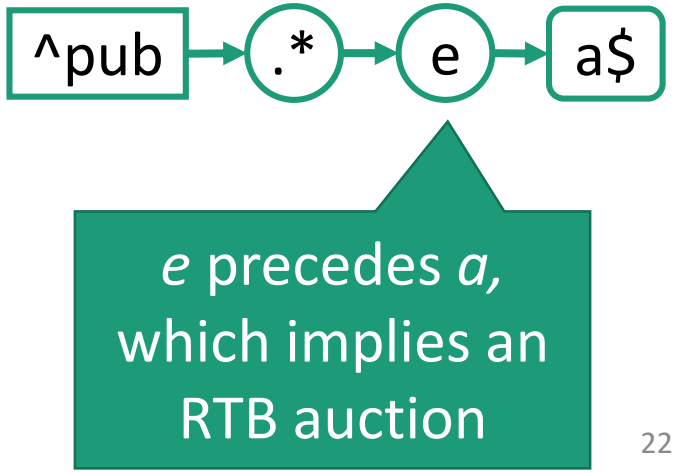
Publisher-side

Example



Rule



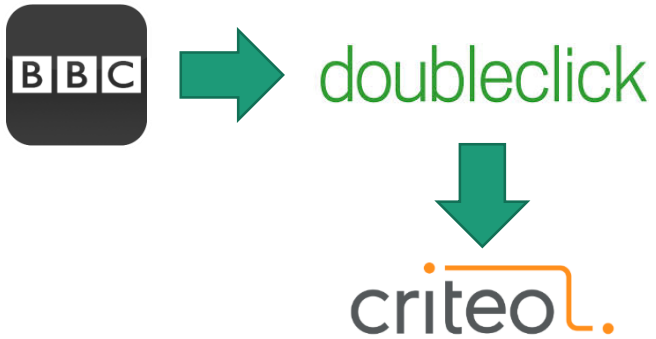


2) Cookie Matching

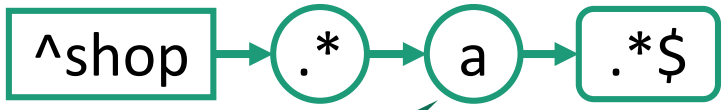
Shopper-side

Publisher-side

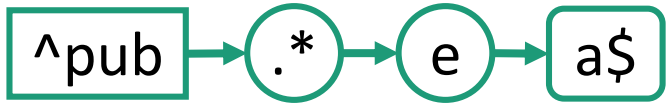
Example



Rule



a must appear on the shopper-side

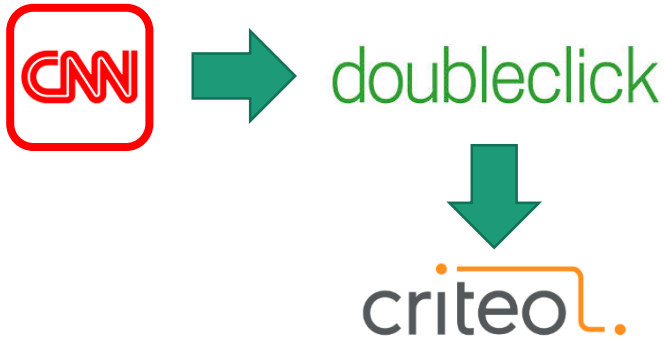


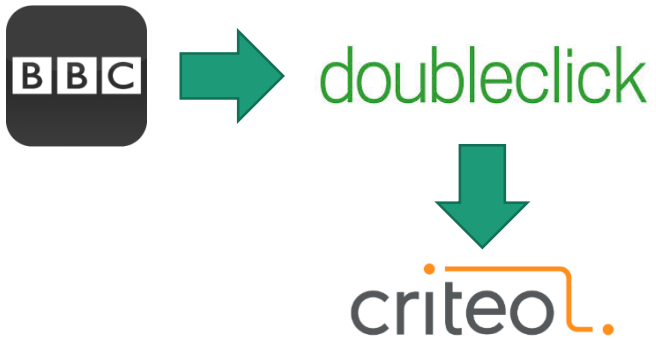
e precedes *a*, which implies an RTB auction

2) Cookie Matching

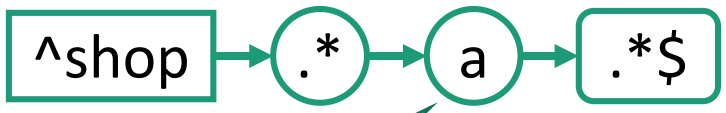
Shopper-side | **Anywhere** | **Publisher-side**

Example

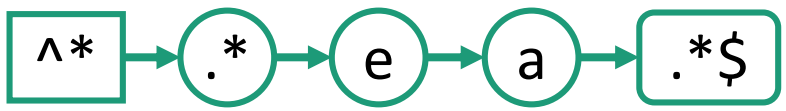




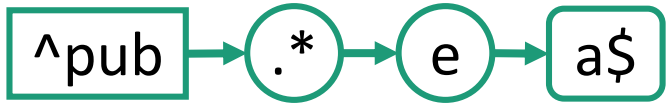
Rule



a must appear on the shopper-side



Transition e→a is where cookie match occurs



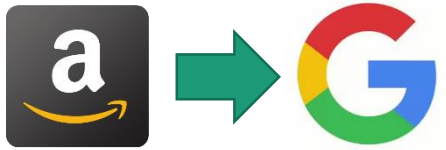
e precedes *a*, which implies an RTB auction

3) Latent (Server-side) Matching

Shopper-side

Publisher-side

Example



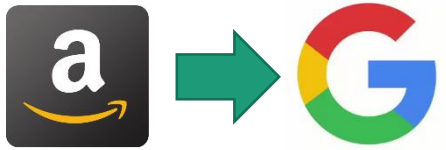
Rule

3) Latent (Server-side) Matching

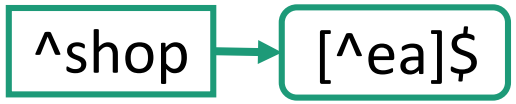
Shopper-side

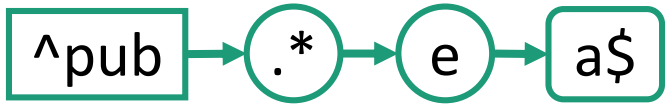
Publisher-side

Example



Rule



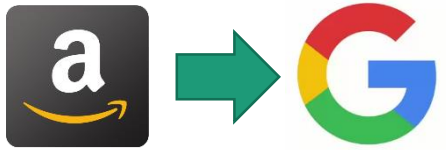


3) Latent (Server-side) Matching

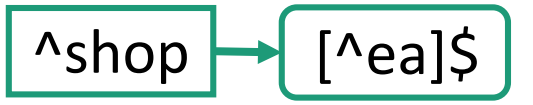
Shopper-side

Publisher-side

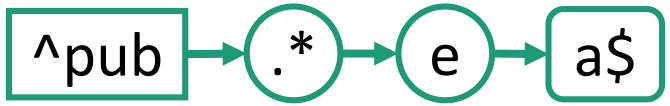
Example



Rule



Neither *e* nor *a* appears on the shopper-side

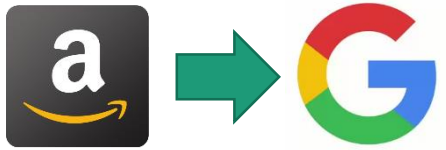


3) Latent (Server-side) Matching

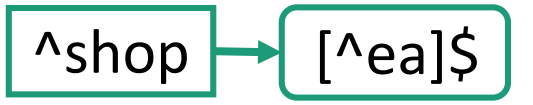
Shopper-side

Publisher-side

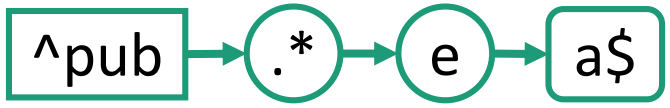
Example



Rule



Neither *e* nor *a* appears on the shopper-side



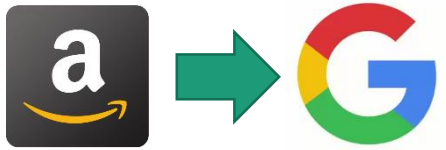
a must receive information from some shopper-side tracker

3) Latent (Server-side) Matching

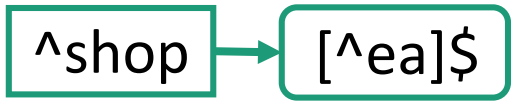
Shopper-side

Publisher-side

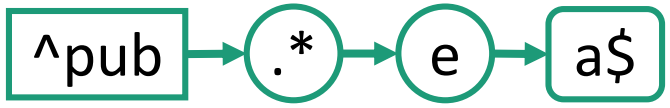
Example



Rule



Neither *e* nor *a* appears on the shopper-side



a must receive information from some shopper-side tracker

We find latent matches in practice!

Data Collection
Classifying Ad Network Flows

Results

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
No Match	**775**	**2**

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
No Match	**775**	**2**

Take away:

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
No Match	**775**	**2**

Take away:

1- As expected, most retargets are due to cookie matching

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
No Match	**775**	**2**

Take away:

1- As expected, most retargets are due to cookie matching

2- Very small number of chains that cannot be categorized

- Suggests low false positive rate of AMT image labeling task

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
No Match	**775**	**2**

Take away:

1- As expected, most retargets are due to cookie matching

2- Very small number of chains that cannot be categorized

- Suggests low false positive rate of AMT image labeling task

3- Surprisingly large amount latent matches…

Categorizing Chains

	Raw Chains		Clustered Chains	
Type	**Chains**	**%**	**Chains**	**%**
Direct (Trivial) Match	1770	5	8449	24
Cookie Match	25049	71	25873	73
Latent (Server-side) Match	5362	15	343	1
No Match	**775**	**2**	**183**	**1**

Cluster together domains by “owner”

- E.g. google.com, doubleclick.com, googlesyndication.com

Categorizing Chains

	Raw Chains		Clustered Chains	
Type	**Chains**	**%**	**Chains**	**%**
Direct (Trivial) Match	1770	5	8449	24
Cookie Match	25049	71	25873	73
Latent (Server-side) Match	5362	15	343	1
No Match	**775**	**2**	**183**	**1**

Cluster together domains by “owner”

- E.g. google.com, doubleclick.com, googlesyndication.com

Latent matches essentially disappear

- The vast majority of these chains involve Google
- Suggests that Google shares tracking data across their services

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	←→	googlesyndication	9090	1887	←→ P
criteo	←→	doubleclick	3610	1144	→ E, P ← DC, P
criteo	←→	adnxs	3263	1066	←→ E, P
criteo	←→	rubiconproject	1586	749	←→ E, P
criteo	←→	servedbyopenx	707	460	←→ P
doubleclick	←→	steelhousemedia	362	27	→ P ← E, P
mathtag	←→	mediaforge	360	124	←→ E, P
netmng	←→	scene7	267	119	→ E ← ?
googlesyndication	←→	adsrvr	107	29	←→ P
rubiconproject	←→	steelhousemedia	86	30	←→ E
googlesyndication	←→	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	←→	adnxs	31	8	?

Heuristics Key
(used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

31% of cookie matching partners would be missed.

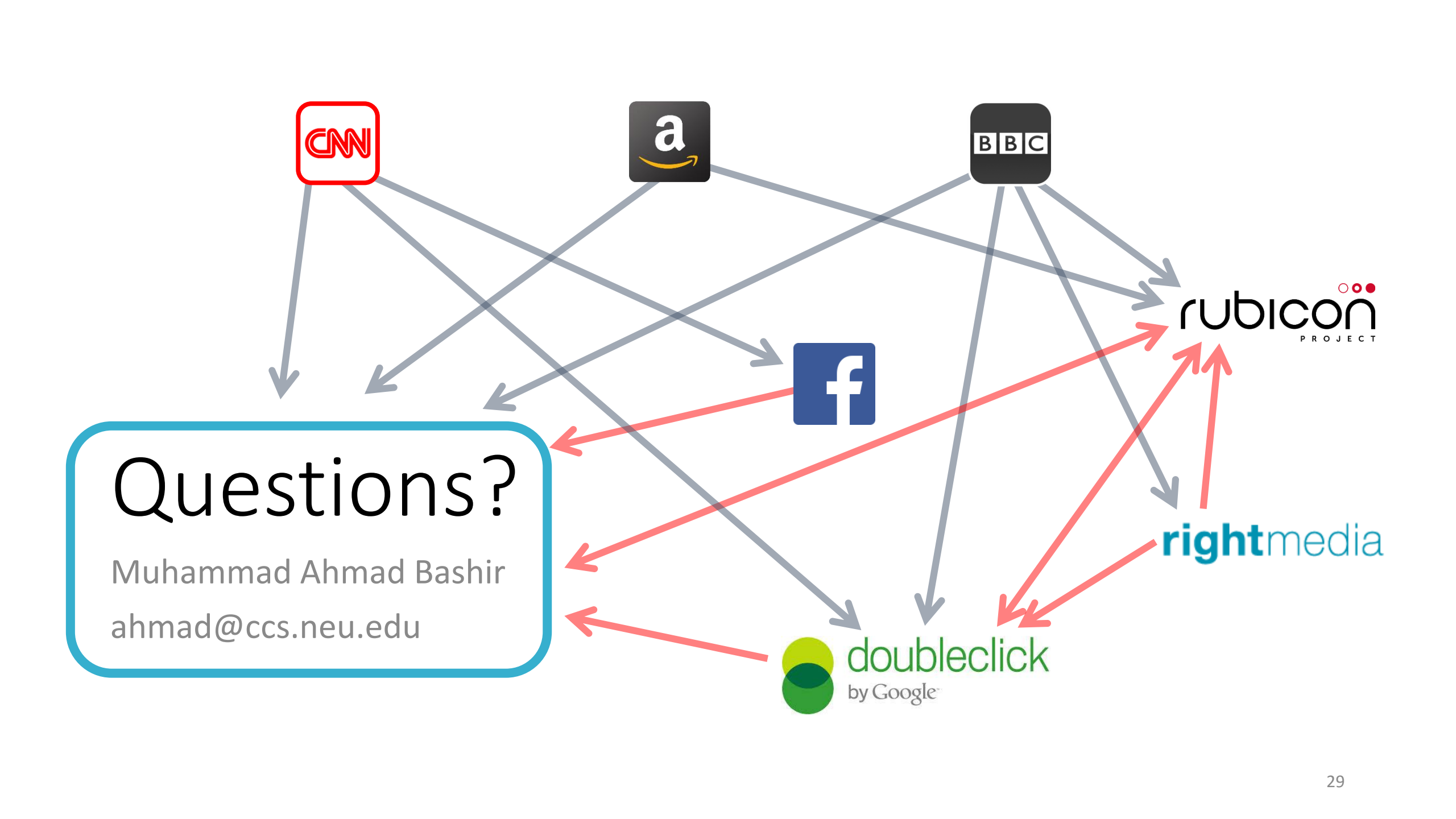
Summary

We develop a novel methodology to detect information flows between ad exchanges

- Controlled methodology enables causal inference
- Defeats obfuscation attempts
- Detects client- and server-side flows

Dataset gives a better picture of ad ecosystem

- Reveals which ad exchanges are linking information about users
- Allows us to reason about how information is being transferred



Questions?

Muhammad Ahmad Bashir

ahmad@ccs.neu.edu

Inclusion Chains

- Instrumented Chromium binary that records the provenance of page elements
  - Uses Information Flow Analysis techniques (IFA)
  - Handles Flash, *exec()*, *setTimeout()*, cross-frame, inline scripts, etc.

Inclusion Chains

- Instrumented Chromium binary that records the provenance of page elements
  - Uses Information Flow Analysis techniques (IFA)
  - Handles Flash, *exec()*, *setTimeout()*, cross-frame, inline scripts, etc.

DOM: a.com/index.html

```
<html>
<body>
  <script src="b.com/adlib.js"></script>
  <iframe src="c.net/adbox.html">
    <html>
      <script src="code.js"></script>
      <object data="d.org/flash.swf">
      </object>
    </html>
  </iframe>
</body>
</html>
```

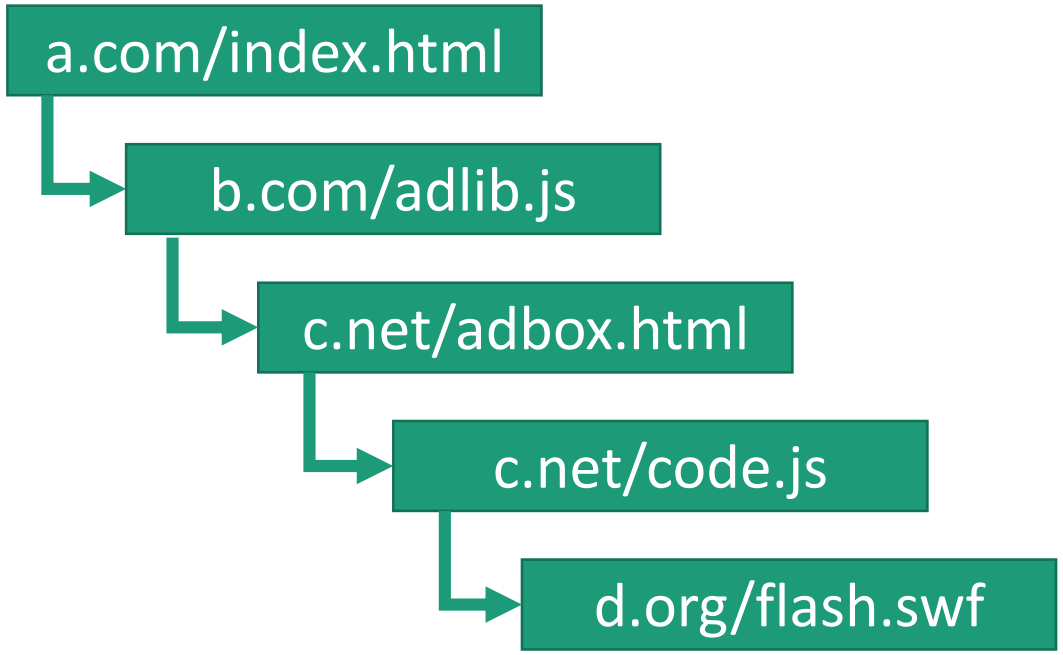
Inclusion Chains

- Instrumented Chromium binary that records the provenance of page elements
  - Uses Information Flow Analysis techniques (IFA)
  - Handles Flash, *exec()*, *setTimeout()*, cross-frame, inline scripts, etc.

DOM: a.com/index.html

```
<html>
<body>
  <script src="b.com/adlib.js"></script>
  <iframe src="c.net/adbox.html">
    <html>
      <script src="code.js"></script>
      <object data="d.org/flash.swf">
      </object>
    </html>
  </iframe>
</body>
</html>
```

Inclusion Chain



3) Indirect Matching

Shopper-side

Publisher-side

Example

Rule

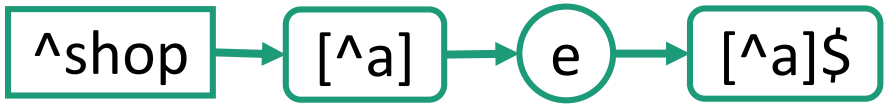
3) Indirect Matching

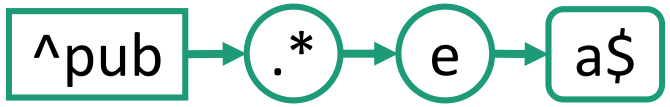
Shopper-side

Publisher-side

Example

Rule





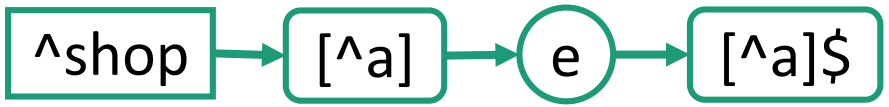
3) Indirect Matching

Shopper-side

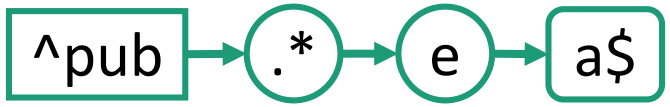
Publisher-side

Example

Rule



Only the exchange e appears on the shopper-side…



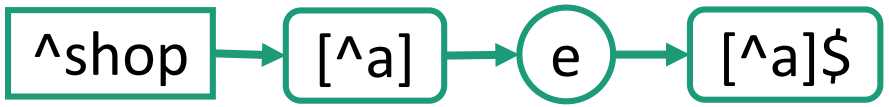
3) Indirect Matching

Shopper-side

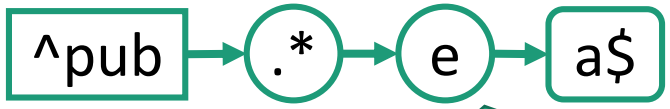
Publisher-side

Example

Rule



Only the exchange e appears on the shopper-side…



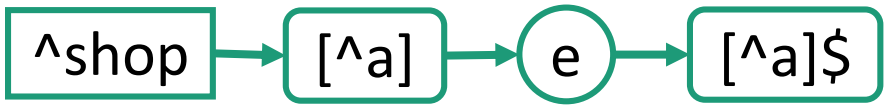
e must pass browsing history data to participants in the auction, thus no cookie matching is necessary

3) Indirect Matching

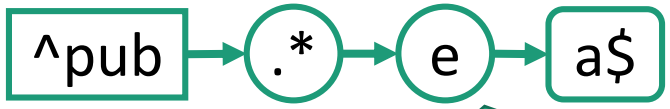
Shopper-side | **Publisher-side**

Example

Rule



Only the exchange e appears on the shopper-side…



e must pass browsing history data to participants in the auction, thus no cookie matching is necessary

We do not expect to find indirect matches in the data.

References

Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz. “The web never forgets: Persistent tracking mechanisms in the wild.” CCS, 2014.

Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, Christo Wilson. “Tracing Information Flows Between Ad Exchanges Using Retargeted Ads.” Usenix Security, 2016.

Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, Richard Mortier. “Tracking personal identifiers across the web.” PAM, 2016.

Lukasz Olejnik, Tran Minh-Dung, Claude Castelluccia. “Selling off privacy at auction.” NDSS, 2014.

Filtering Images

Filter	Total Unique Images
All images from the crawlers	571,636

Filtering Images

Filter	Total Unique Images
All images from the crawlers	571,636
Use EasyList to identify advertisements	93,726

Filtering Images

Filter	Total Unique Images
All images from the crawlers	571,636
Use EasyList to identify advertisements	93,726
Remove ads that are shown to >1 persona	31,850

- Personas visited non-overlapping retailers
  - By definition, retargets should only be shown to a single persona

Filtering Images

Filter	Total Unique Images
All images from the crawlers	571,636
Use EasyList to identify advertisements	93,726
Remove ads that are shown to >1 persona	31,850
Use crowdsourcing to locate retargets	5,102

- Personas visited non-overlapping retailers
  - By definition, retargets should only be shown to a single persona
- Spent $415 uploading 1,142 HITs to Amazon Mechanical Turk
  - Each HIT asked the worker to label 30 ad images
  - 27 were unlabeled, 3 were known retargets (control images)
  - All ads were labeled by 2 workers
  - Any ad identified as *targeted* was also manually inspected by us