MEASUREMENT-DRIVEN AUDITING OF REAL-WORLD SYSTEMS

Alan Mislove

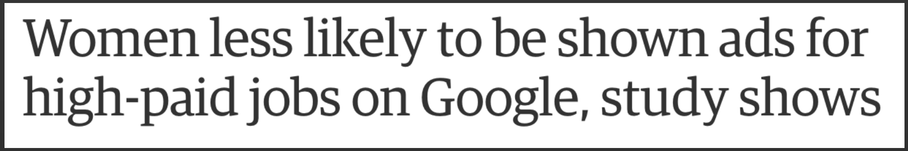
Northeastern University

BIG DATA + ALGORITHMS

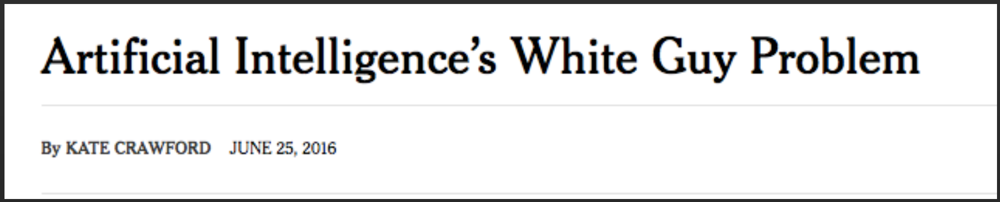
Algorithms driven by big data are beginning to shape our world

- In many cases, these systems are provide useful benefits

However, they may be detrimental to some users

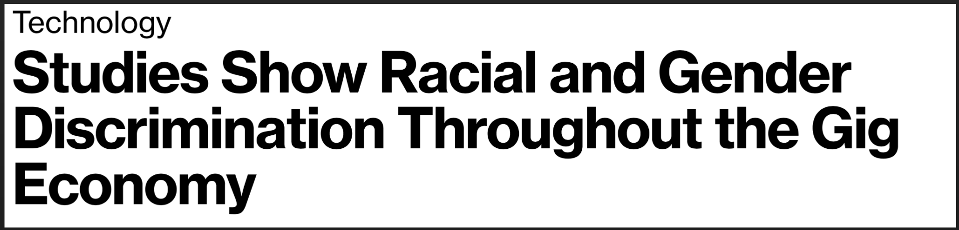


Women less likely to be shown ads for high-paid jobs on Google, study shows



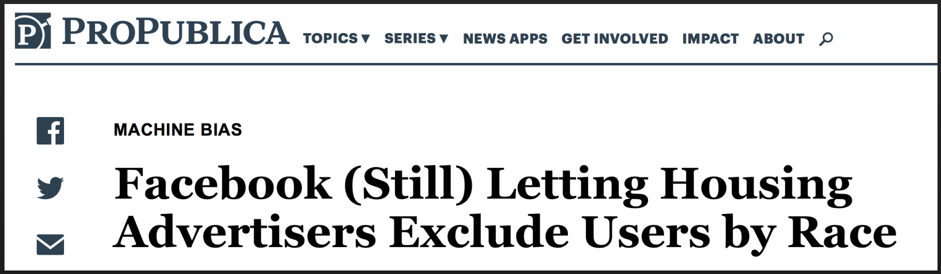
Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016



Technology

Studies Show Racial and Gender Discrimination Throughout the Gig Economy



MACHINE BIAS

Facebook (Still) Letting Housing Advertisers Exclude Users by Race

MEASUREMENT + FAIRNESS

Grew out of systems/networks/measurement research communities

- IMC, IEEE S&P, CCS, WWW, …

Answer questions by increasing transparency of online systems:

- *What data are being used as input to real-world algorithms?*
- *Can we explain some of their output?*
- *Are these systems having detrimental effects on users?*
- *Do these systems have unique vulnerabilities or weaknesses?*

From perspective of outsider (no privileged access)

RESEARCH CHALLENGES

Numerous challenges to studying these systems and algorithms:

① Systems are proprietary black boxes

② Input data is numerous, unknown, and often privacy-sensitive

③ Providers typically somewhat adversarial

Must develop techniques to measure systems from outside

Address ethical concerns, sampling bias, etc…

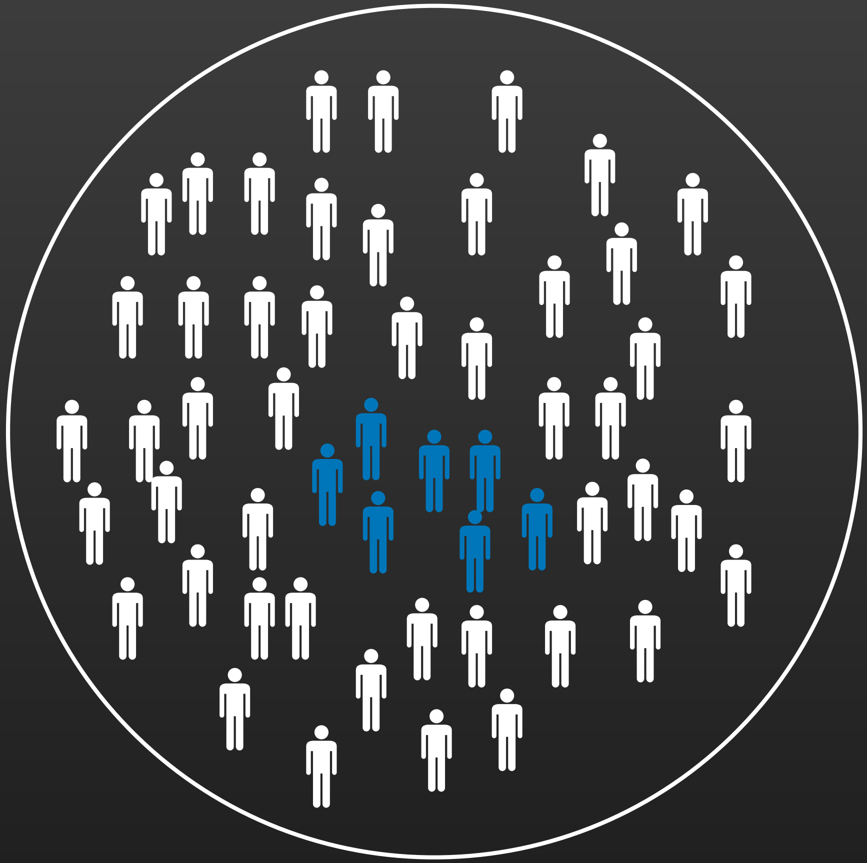
Remainder of talk: Two examples of measuring real-world systems

Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface

[IEEE S&P’18]

THE RISE OF TARGETED ADVERTISING

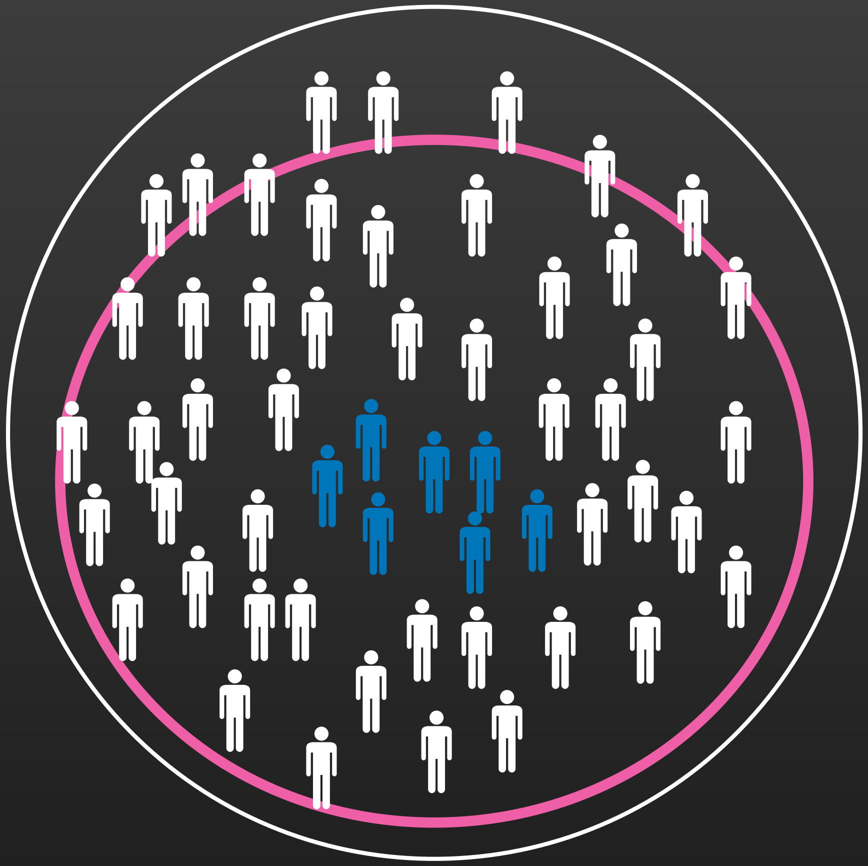
Ad: Musical for teens



THE RISE OF TARGETED ADVERTISING

Ad: Musical for teens

Ad on T.V.



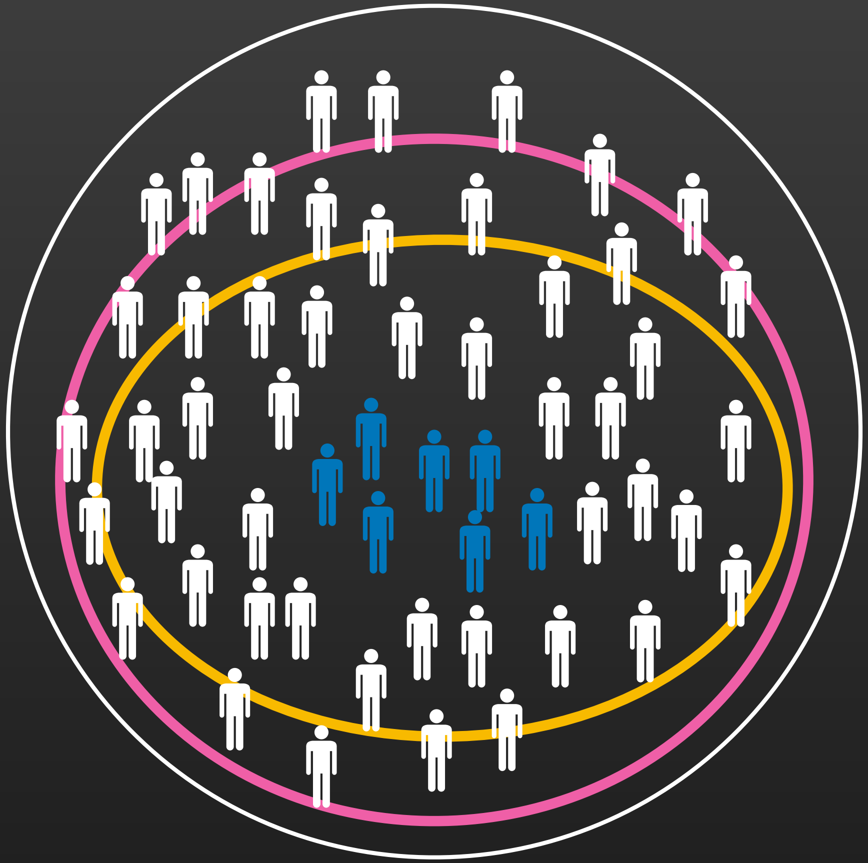
THE RISE OF TARGETED ADVERTISING

Ad: Musical for teens

Ad on T.V.

Ad targeting search keywords

Music, theater



THE RISE OF TARGETED ADVERTISING

Ad: Musical for teens

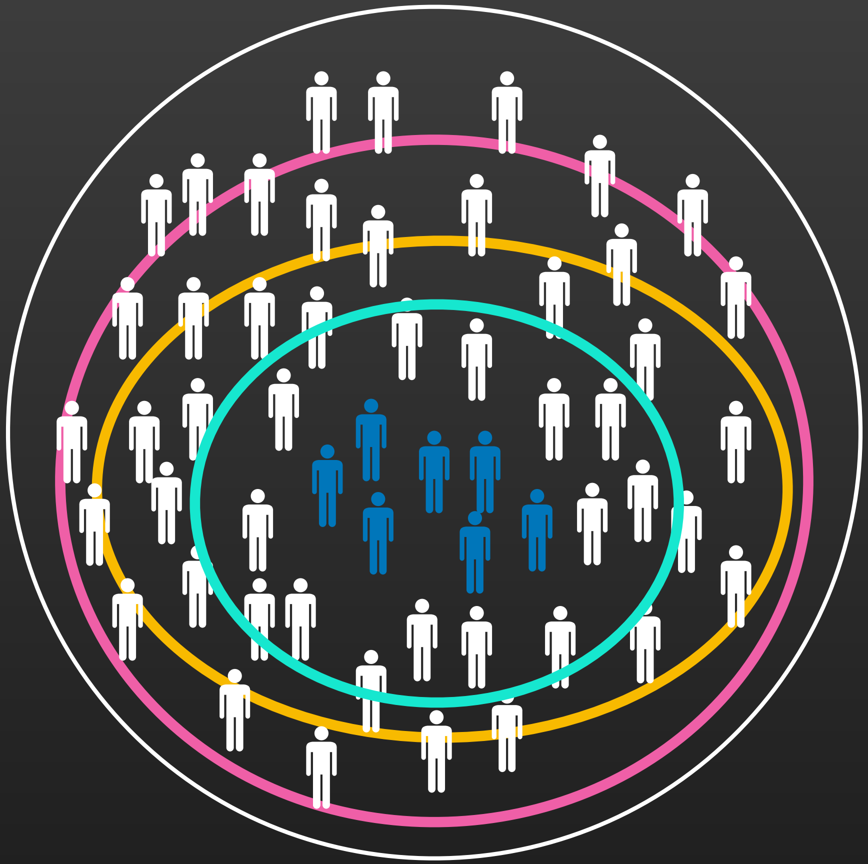
Ad on T.V.

Ad targeting search keywords

Music, theater

Ad targeting user attributes

Teens interested in music and theatre



THE RISE OF TARGETED ADVERTISING

Ad: Musical for teens

Ad on T.V.

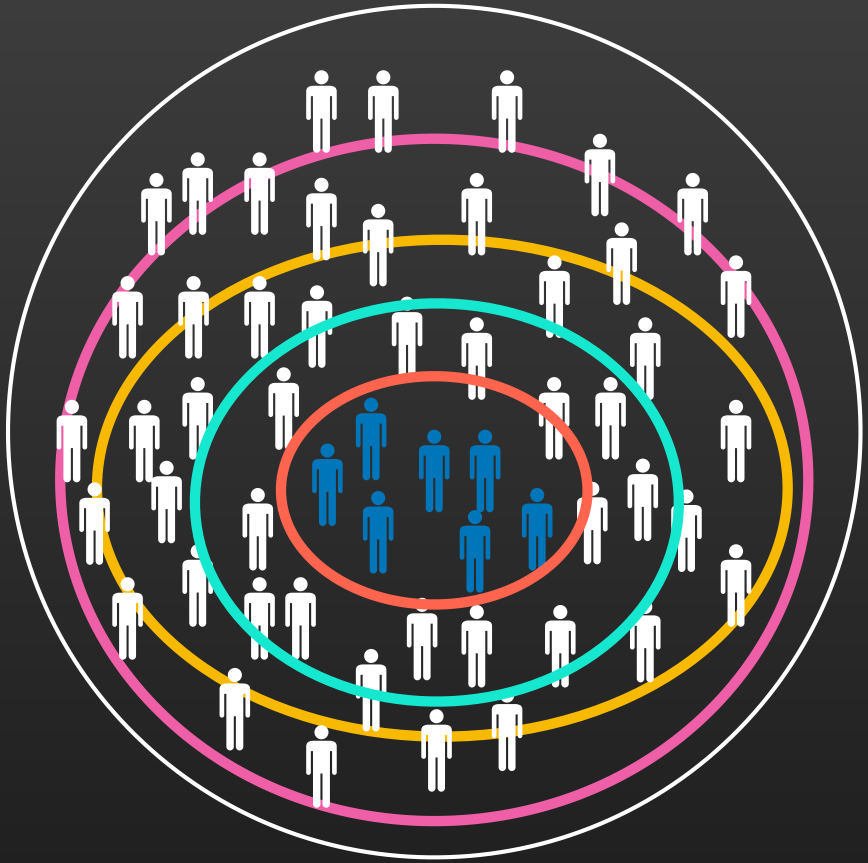
Ad targeting search keywords

Music, theater

Ad targeting user attributes

Teens interested in music and theatre

Ad targeting specific customers?



THE RISE OF TARGETED ADVERTISING

Ad: Musical for teens

Ad on T.V.

Ad targeting search keywords

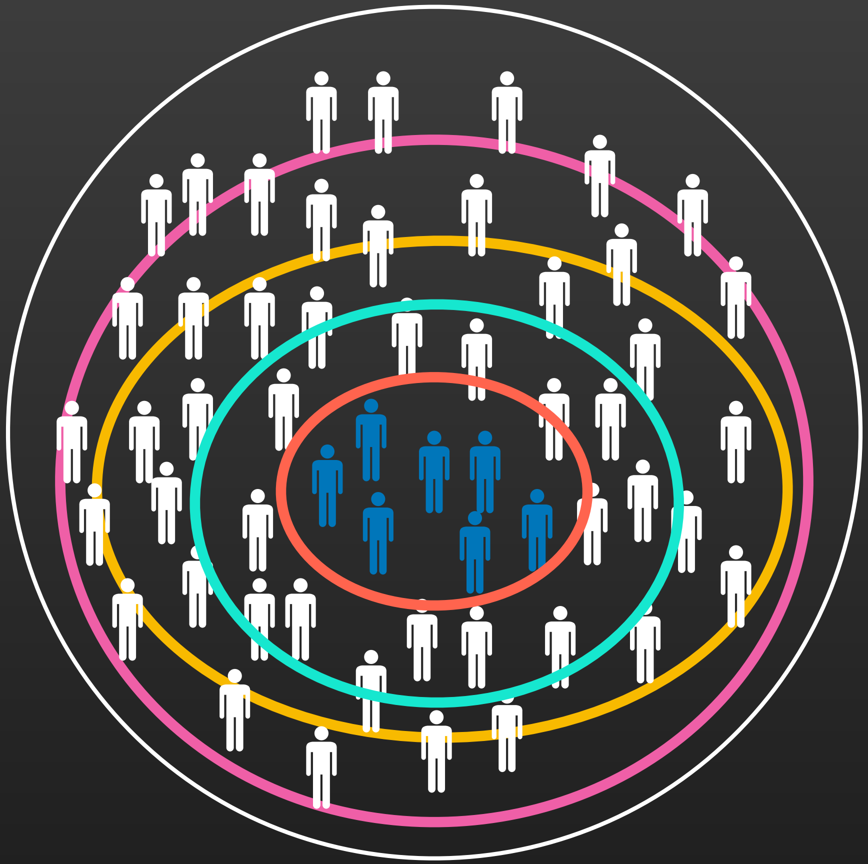
Music, theater

Ad targeting user attributes

Teens interested in music and theatre

Ad targeting specific customers?

Called PII-based targeting !

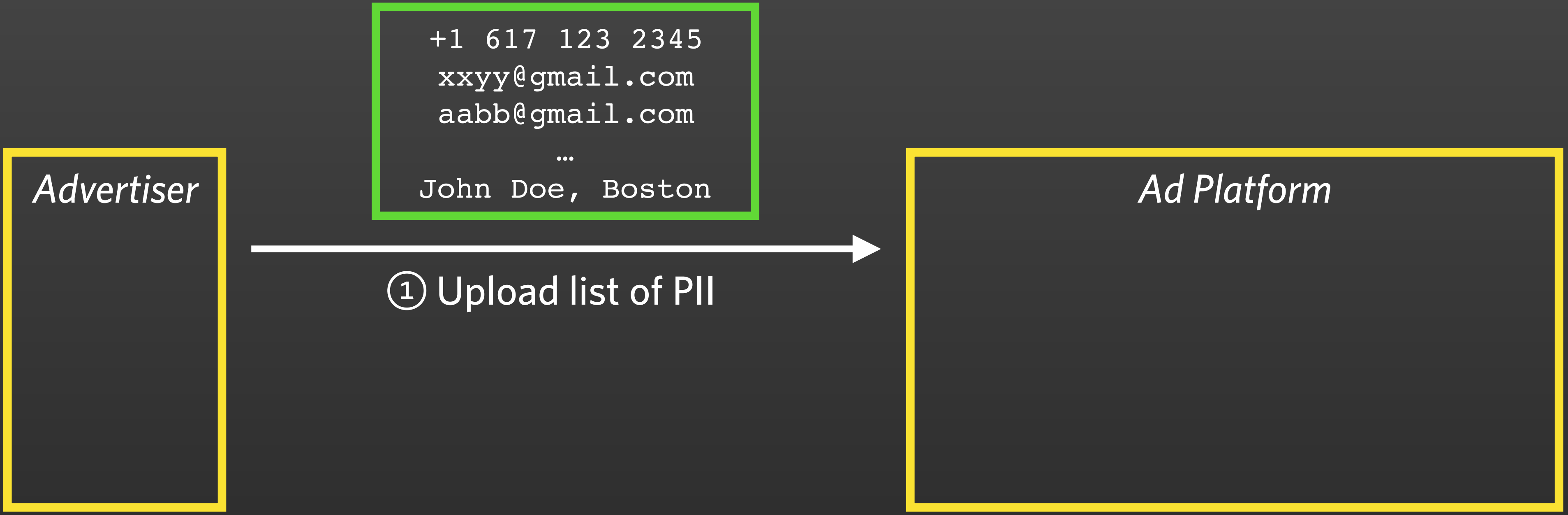


PII-BASED TARGETING ALREADY COMMON

Advertiser

Ad Platform

PII-BASED TARGETING ALREADY COMMON



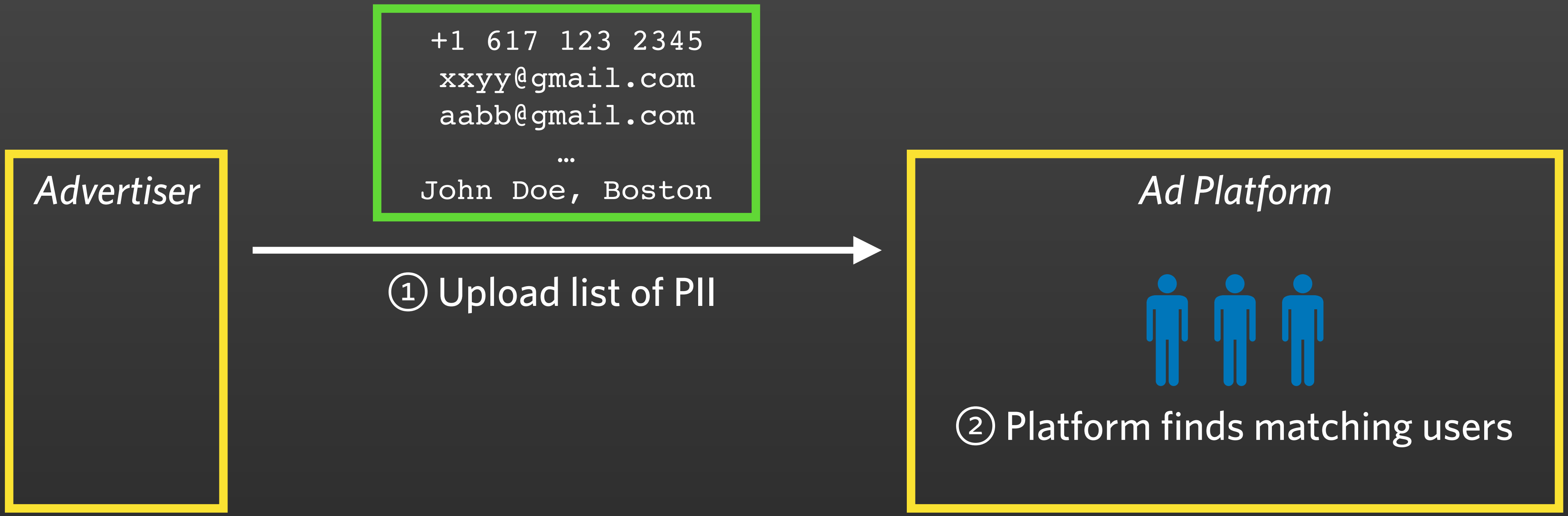
Advertiser

+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston

① Upload list of PII

Ad Platform

PII-BASED TARGETING ALREADY COMMON



Advertiser

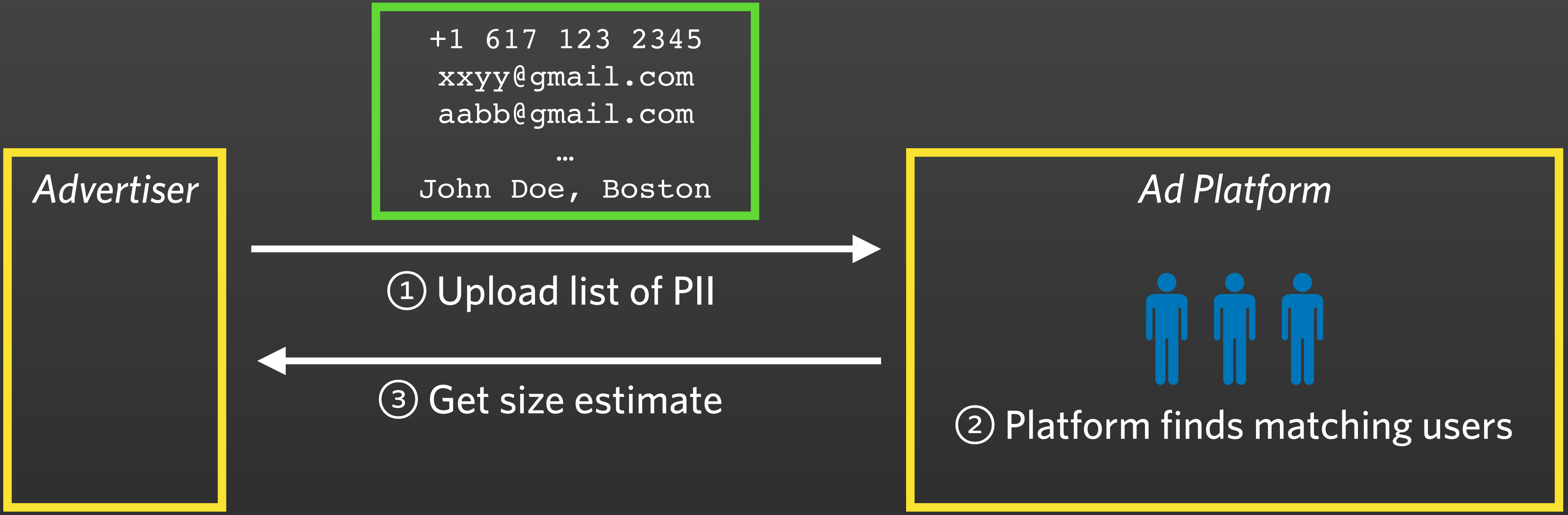
```
+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston
```

① Upload list of PII

Ad Platform

② Platform finds matching users

PII-BASED TARGETING ALREADY COMMON



```
+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston
```

Advertiser

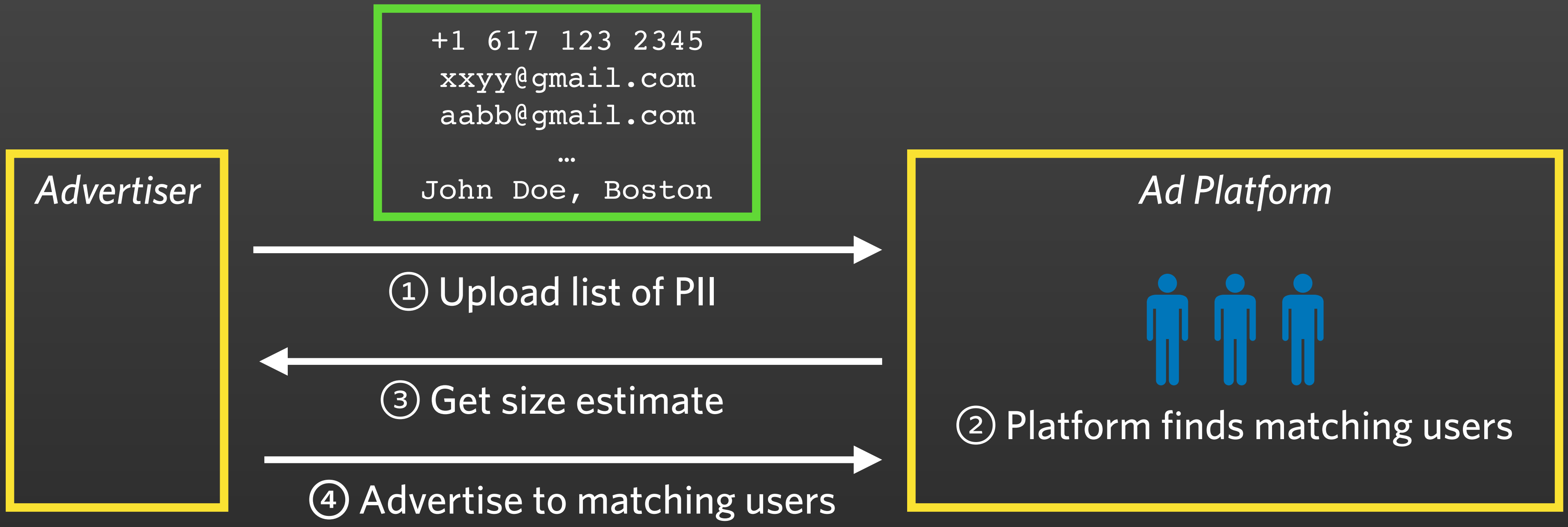
① Upload list of PII

③ Get size estimate

Ad Platform

② Platform finds matching users

PII-BASED TARGETING ALREADY COMMON



```
+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston
```

Advertiser

Ad Platform

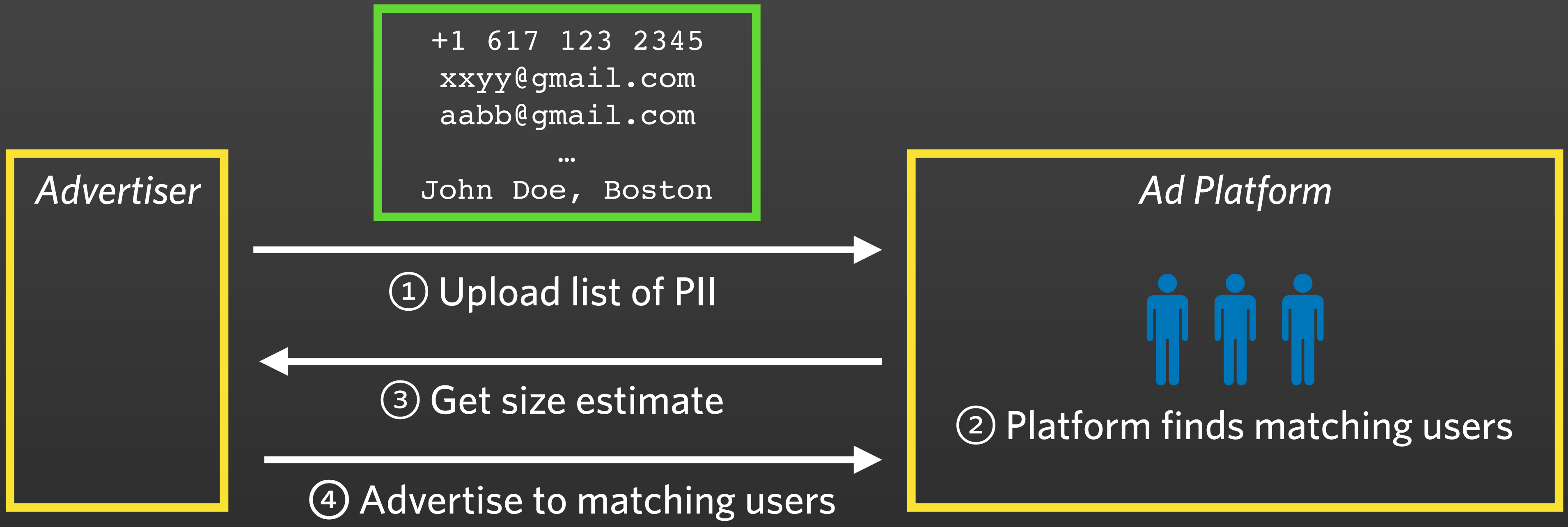
① Upload list of PII

② Platform finds matching users

③ Get size estimate

④ Advertise to matching users

PII-BASED TARGETING ALREADY COMMON



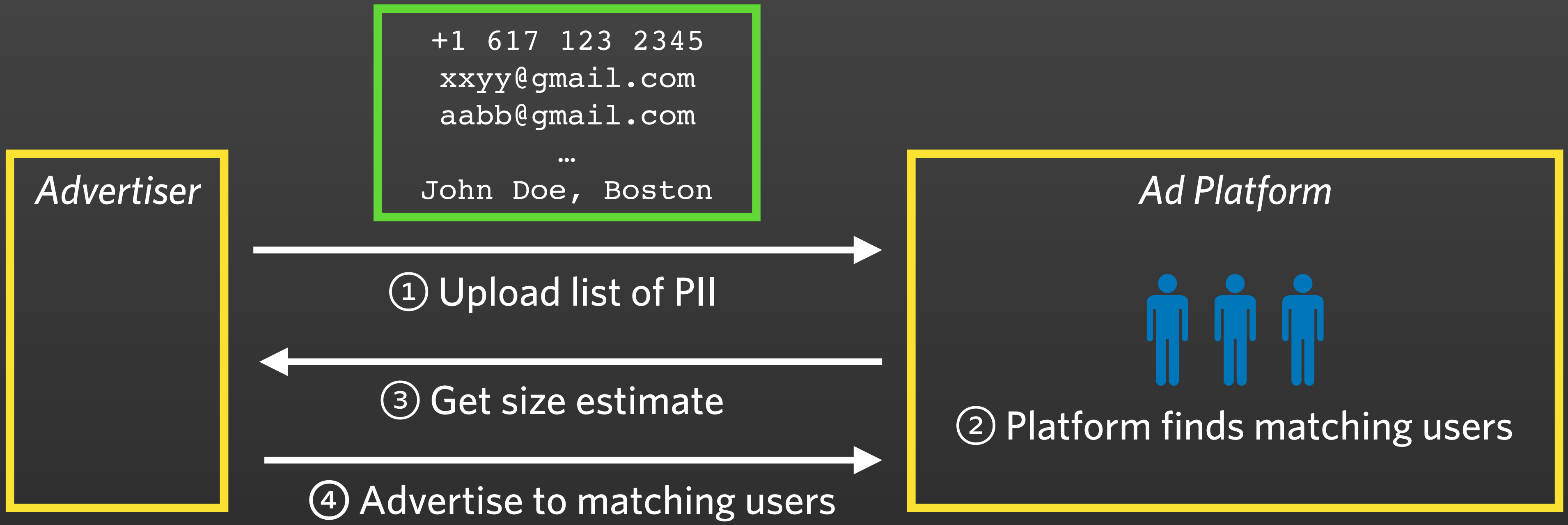
```
+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston
```

Advertiser
Ad Platform
① Upload list of PII
② Platform finds matching users
③ Get size estimate
④ Advertise to matching users

Advantages to advertiser:

1. Pay only for users you want to reach
2. Exploit different external data sources

PII-BASED TARGETING ALREADY COMMON



+1 617 123 2345
xxyy@gmail.com
aabb@gmail.com
…
John Doe, Boston

Advertiser

Ad Platform

① Upload list of PII

② Platform finds matching users

③ Get size estimate

④ Advertise to matching users

Advantages to advertiser:

1. Pay only for users you want to reach
2. Exploit different external data sources

facebook business Custom audiences

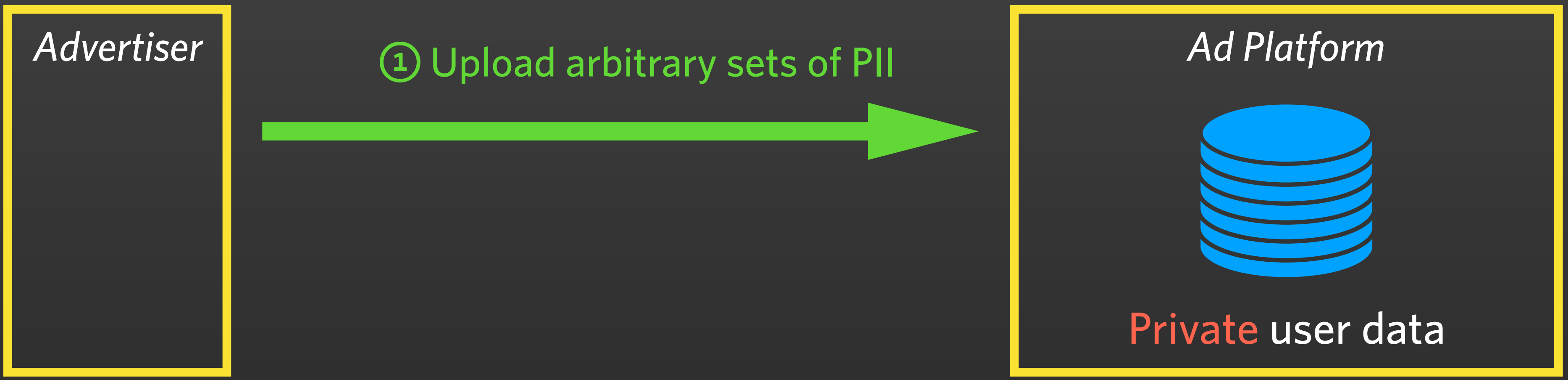
Google AdWords Customer match

Business Tailored audiences

RISKS OF PII-BASED TARGETING

Advertiser
Ad Platform
Private user data

RISKS OF PII-BASED TARGETING



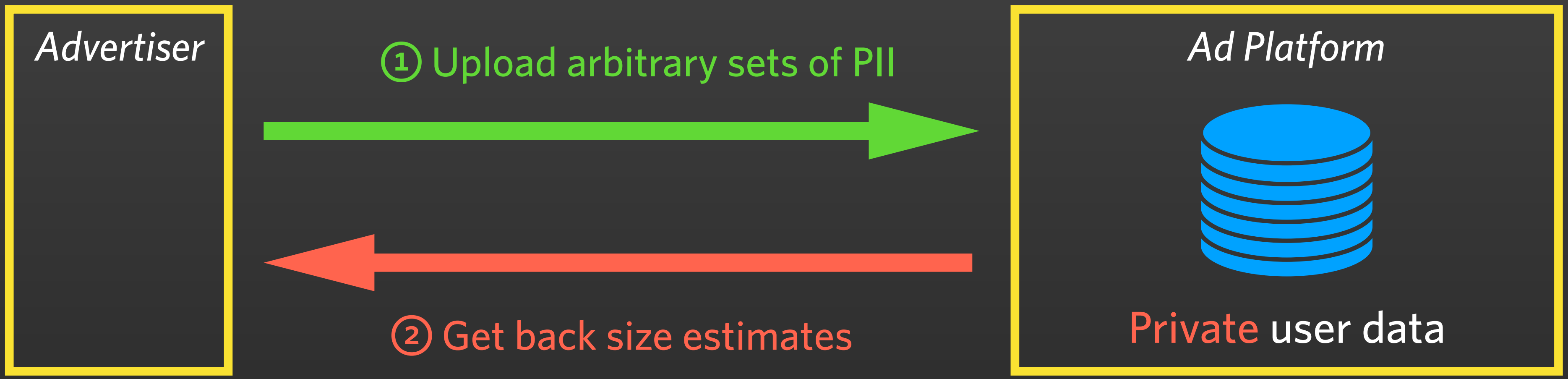
Advertiser

① Upload arbitrary sets of PII

Ad Platform

Private user data

RISKS OF PII-BASED TARGETING



Advertiser

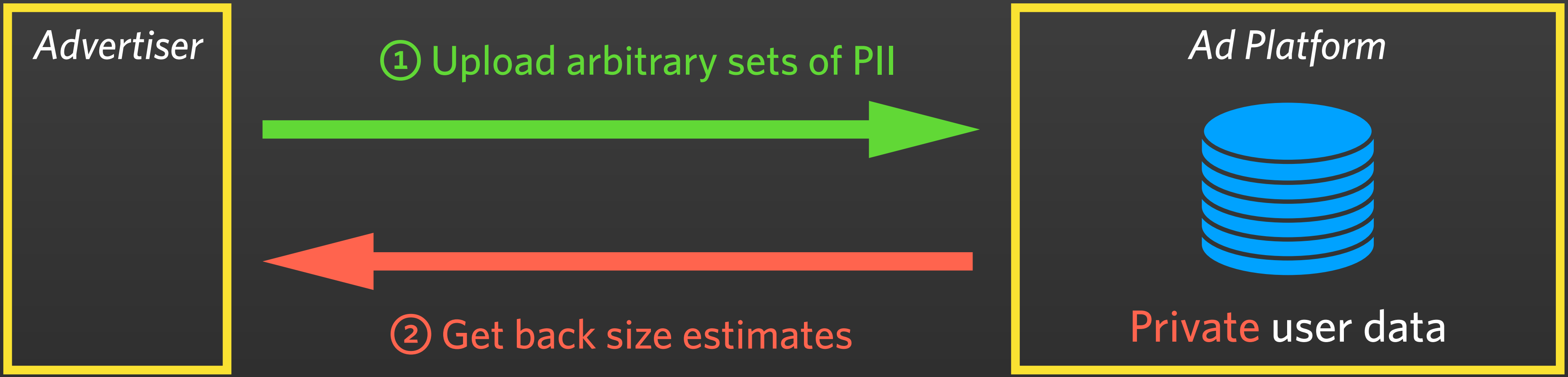
① Upload arbitrary sets of PII

② Get back size estimates

Ad Platform

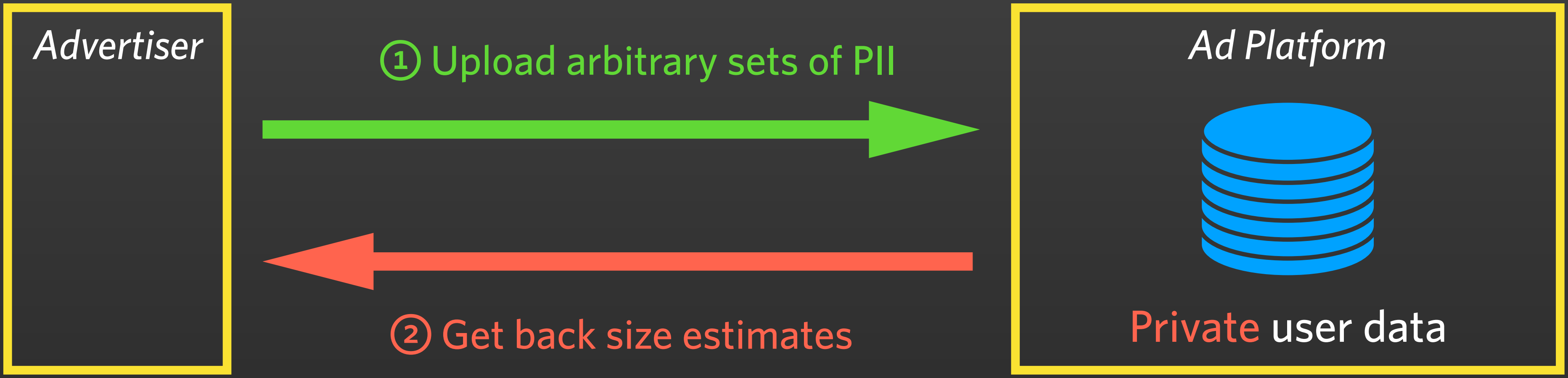
Private user data

RISKS OF PII-BASED TARGETING



This is a query to the user database!

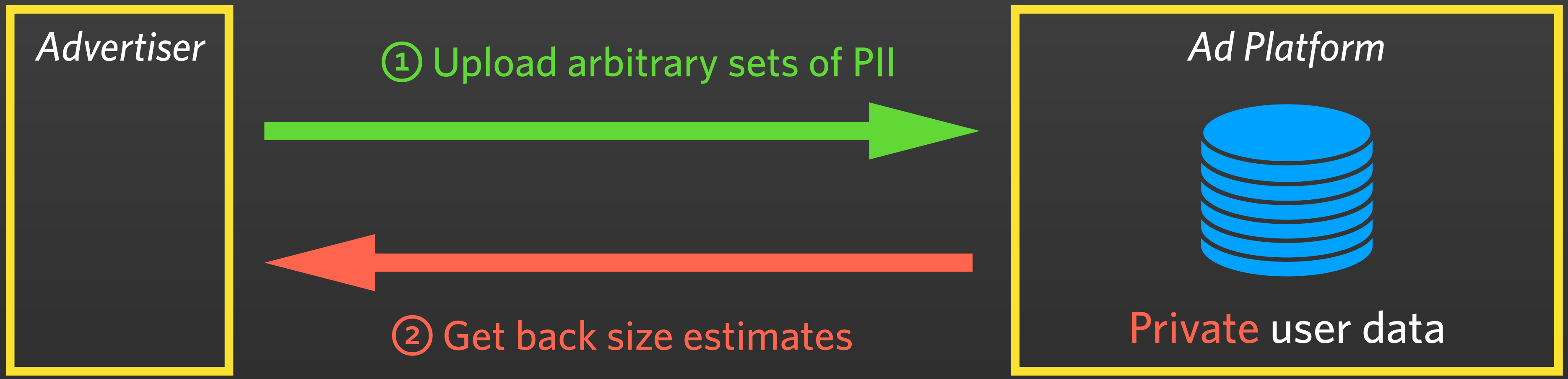
RISKS OF PII-BASED TARGETING



This is a query to the user database!

Could these statistics inadvertently leak user information?

RISKS OF PII-BASED TARGETING



Advertiser

① Upload arbitrary sets of PII

② Get back size estimates

Ad Platform

Private user data

This is a query to the user database!

Could these statistics inadvertently leak user information?

Anyone can be an advertiser…

FEATURES OF FACEBOOK'S SIZE ESTIMATES

1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

FEATURES OF FACEBOOK'S SIZE ESTIMATES

1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

2. Records matching same user are de-duplicated

De-duplication

FEATURES OF FACEBOOK'S SIZE ESTIMATES

1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

2. Records matching same user are de-duplicated

De-duplication

```
xxxyyyzzz@gmail.com

aaabbbccc@gmail.com

  +1 617 888 9999
```

Upload data

FEATURES OF FACEBOOK'S SIZE ESTIMATES

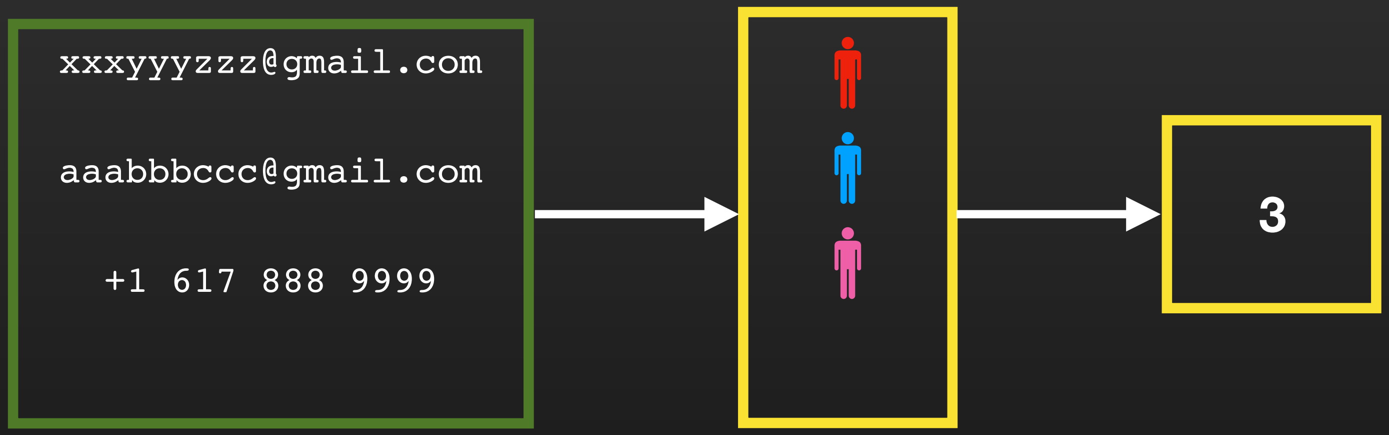
1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

2. Records matching same user are de-duplicated

De-duplication



xxxyyyzzz@gmail.com

aaabbbccc@gmail.com

+1 617 888 9999

Upload data

Facebook calculates size estimate

FEATURES OF FACEBOOK'S SIZE ESTIMATES

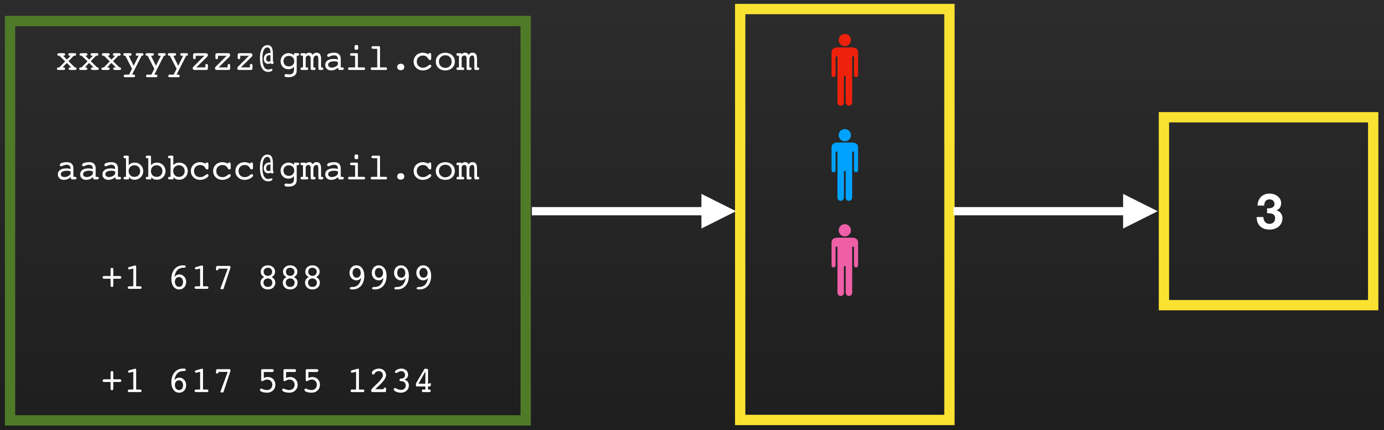
1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

2. Records matching same user are de-duplicated

De-duplication



xxxyyyzzz@gmail.com

aaabbbccc@gmail.com

+1 617 888 9999

+1 617 555 1234

Upload data

Facebook calculates size estimate

FEATURES OF FACEBOOK'S SIZE ESTIMATES

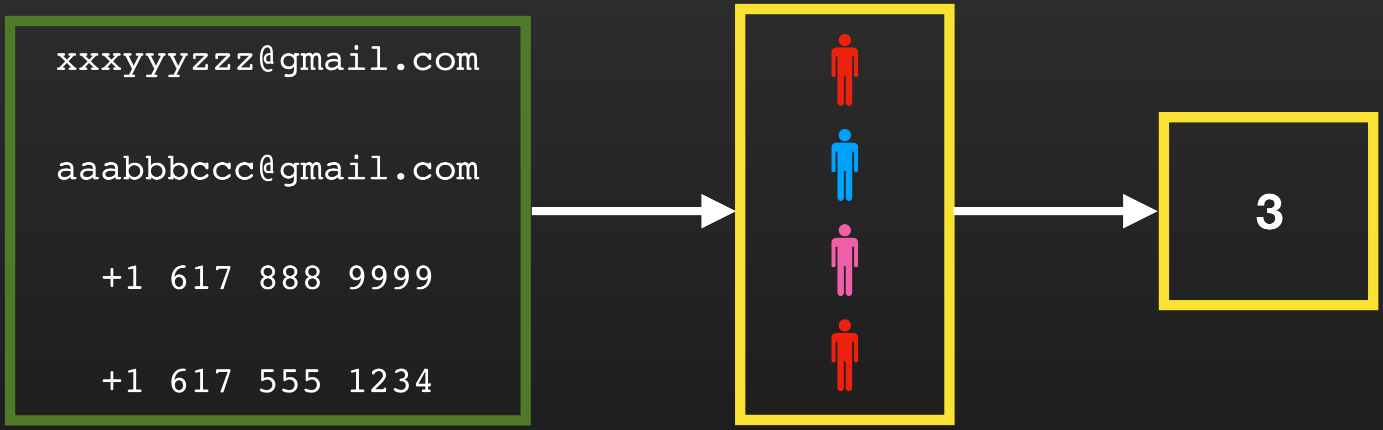
1. Size estimates obfuscated by simple rounding

Obfuscation

{20, 30, 40, ..., 1000, 1100, 1200, ..., 10000, 11000, 12000, ... }

2. Records matching same user are de-duplicated

De-duplication



xxxyyyzzz@gmail.com

aaabbbccc@gmail.com

+1 617 888 9999

+1 617 555 1234

3

Upload data

Facebook calculates size estimate

EXPLOITING THESE FEATURES

EXPLOITING THESE FEATURES

```
+1 617 335 1234
+1 617 111 1534
        …
+1 617 677 9876
```

EXPLOITING THESE FEATURES

Victim’s email

```
+1 617 335 1234
+1 617 111 1534
       …
+1 617 677 9876
```

EXPLOITING THESE FEATURES

Is Victim’s email in

```
+1 617 335 1234
+1 617 111 1534
       …
+1 617 677 9876
```

?

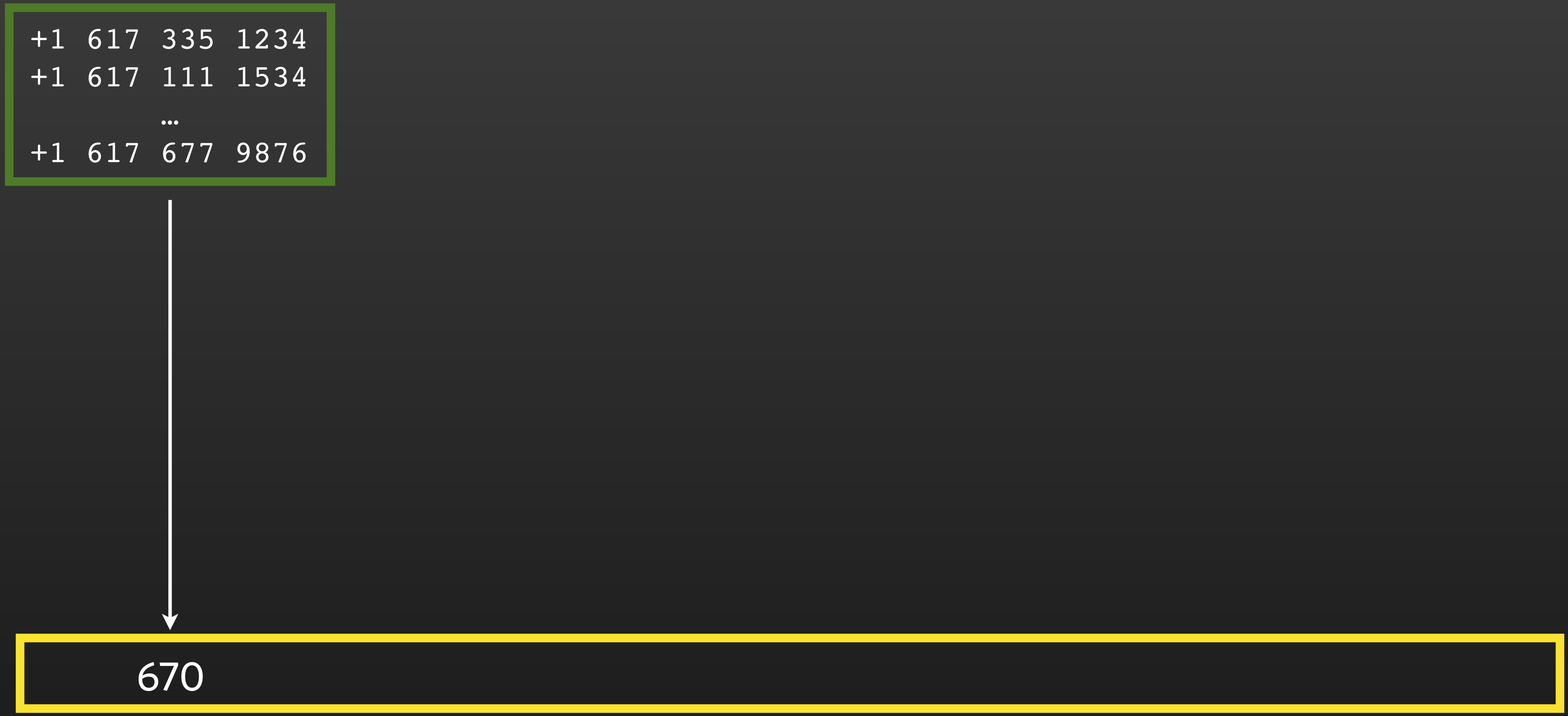
EXPLOITING THESE FEATURES

Goal: Given victim’s email address, find if victim is in a given list of phone numbers

```
+1 617 335 1234
+1 617 111 1534
         …
+1 617 677 9876
```

EXPLOITING THESE FEATURES

Goal: Given victim’s email address, find if victim is in a given list of phone numbers



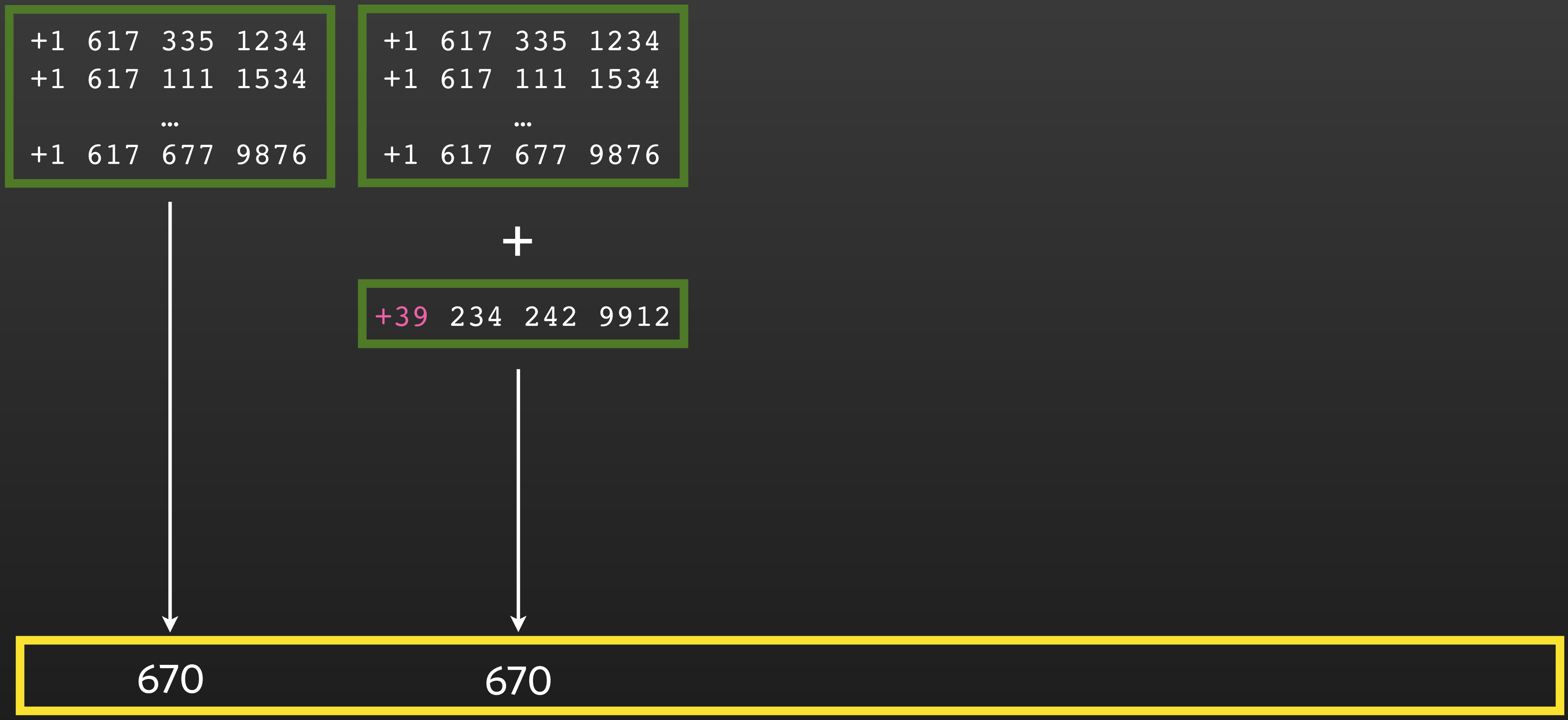
```
+1 617 335 1234
+1 617 111 1534
        …
+1 617 677 9876
```

670

Size estimate

EXPLOITING THESE FEATURES

Goal: Given victim's email address, find if victim is in a given list of phone numbers



```
+1 617 335 1234
+1 617 111 1534
       …
+1 617 677 9876
```

```
+1 617 335 1234
+1 617 111 1534
       …
+1 617 677 9876
```

+

```
+39 234 242 9912
```

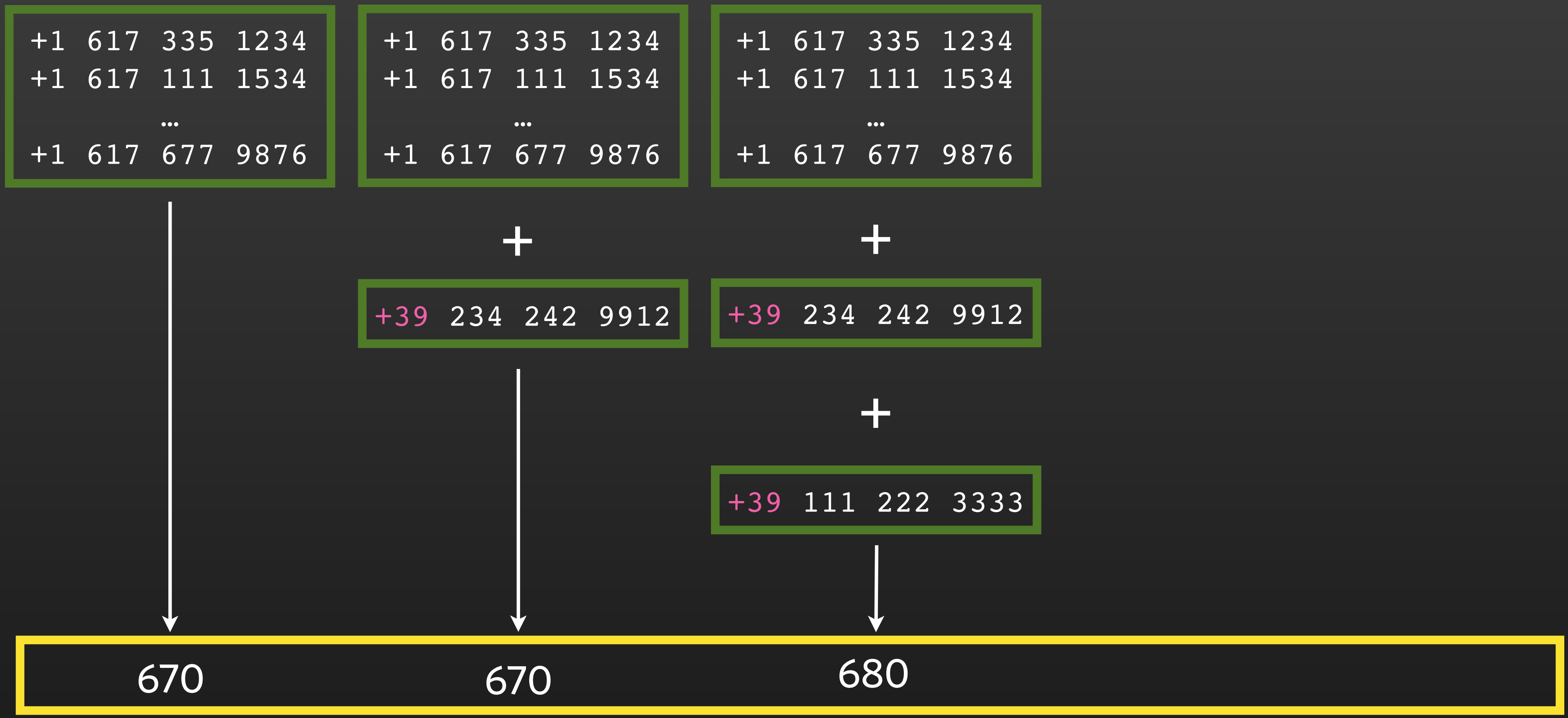
670

670

Size estimate

EXPLOITING THESE FEATURES

Goal: Given victim’s email address, find if victim is in a given list of phone numbers



+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+

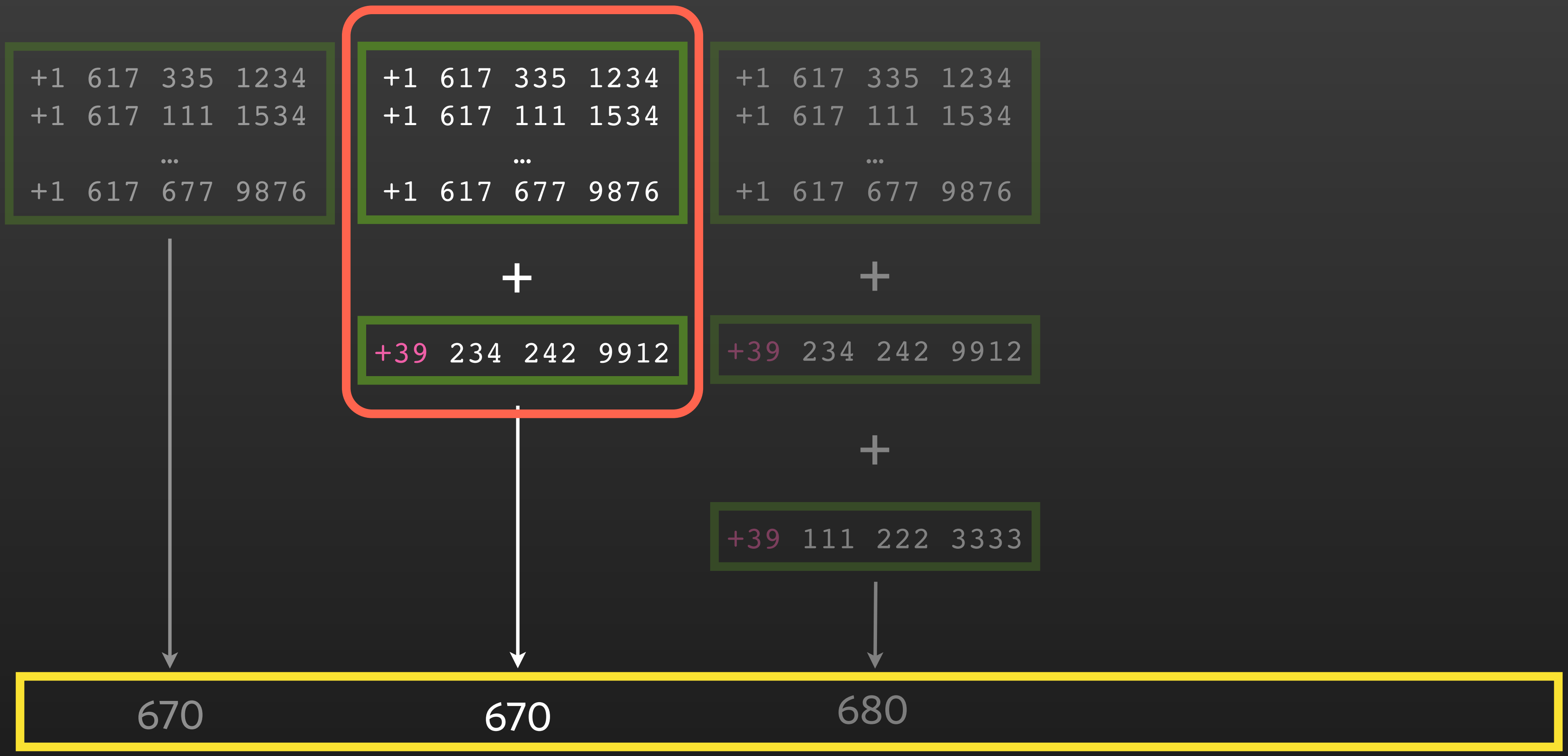
+39 111 222 3333

670 670 680

Size estimate

EXPLOITING THESE FEATURES

Goal: Given victim's email address, find if victim is in a given list of phone numbers



+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+

+39 111 222 3333

670

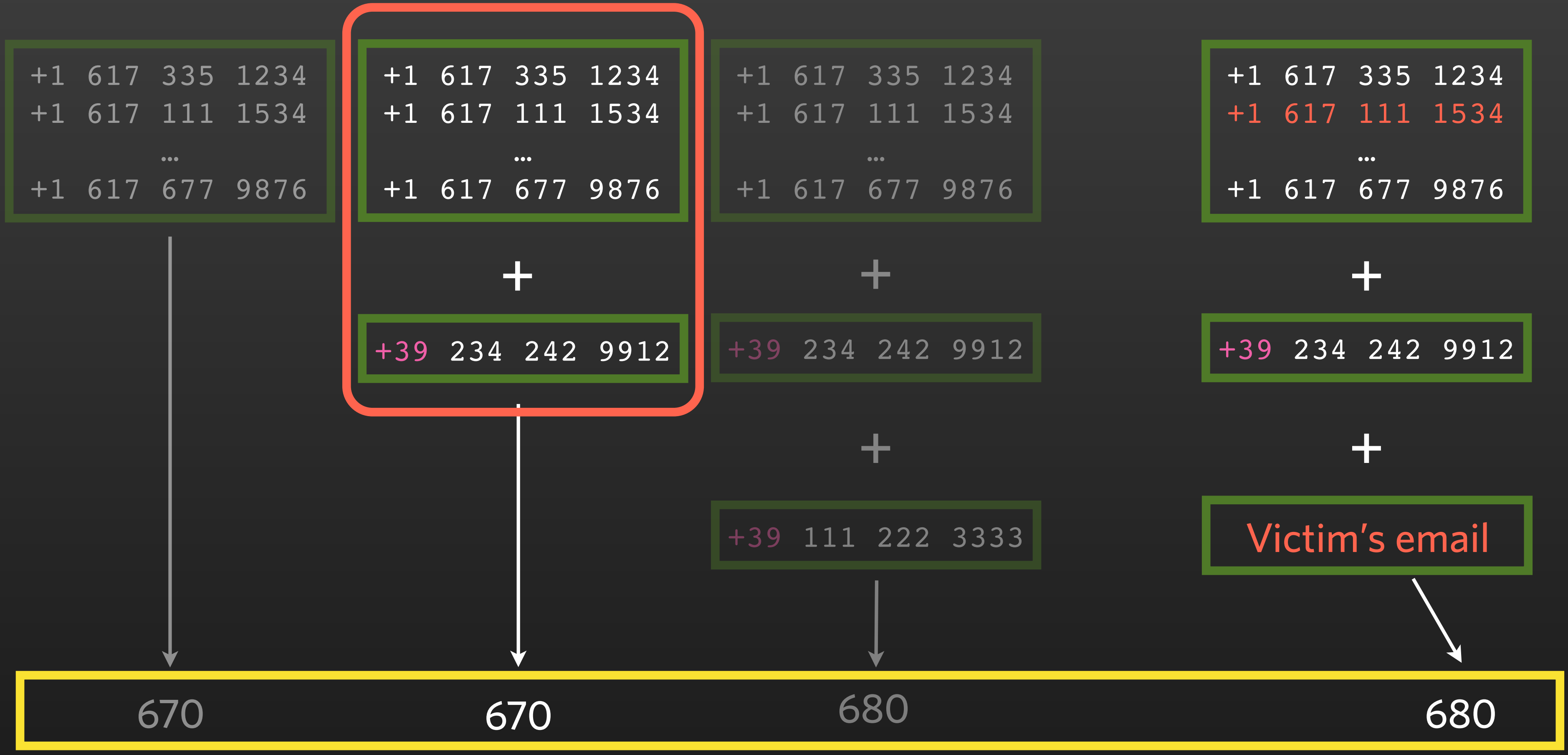
670

680

Size estimate

EXPLOITING THESE FEATURES

Goal: Given victim's email address, find if victim is in a given list of phone numbers



+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

670

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

670

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+

+39 111 222 3333

680

+1 617 335 1234
+1 617 111 1534
…
+1 617 677 9876

+

+39 234 242 9912

+

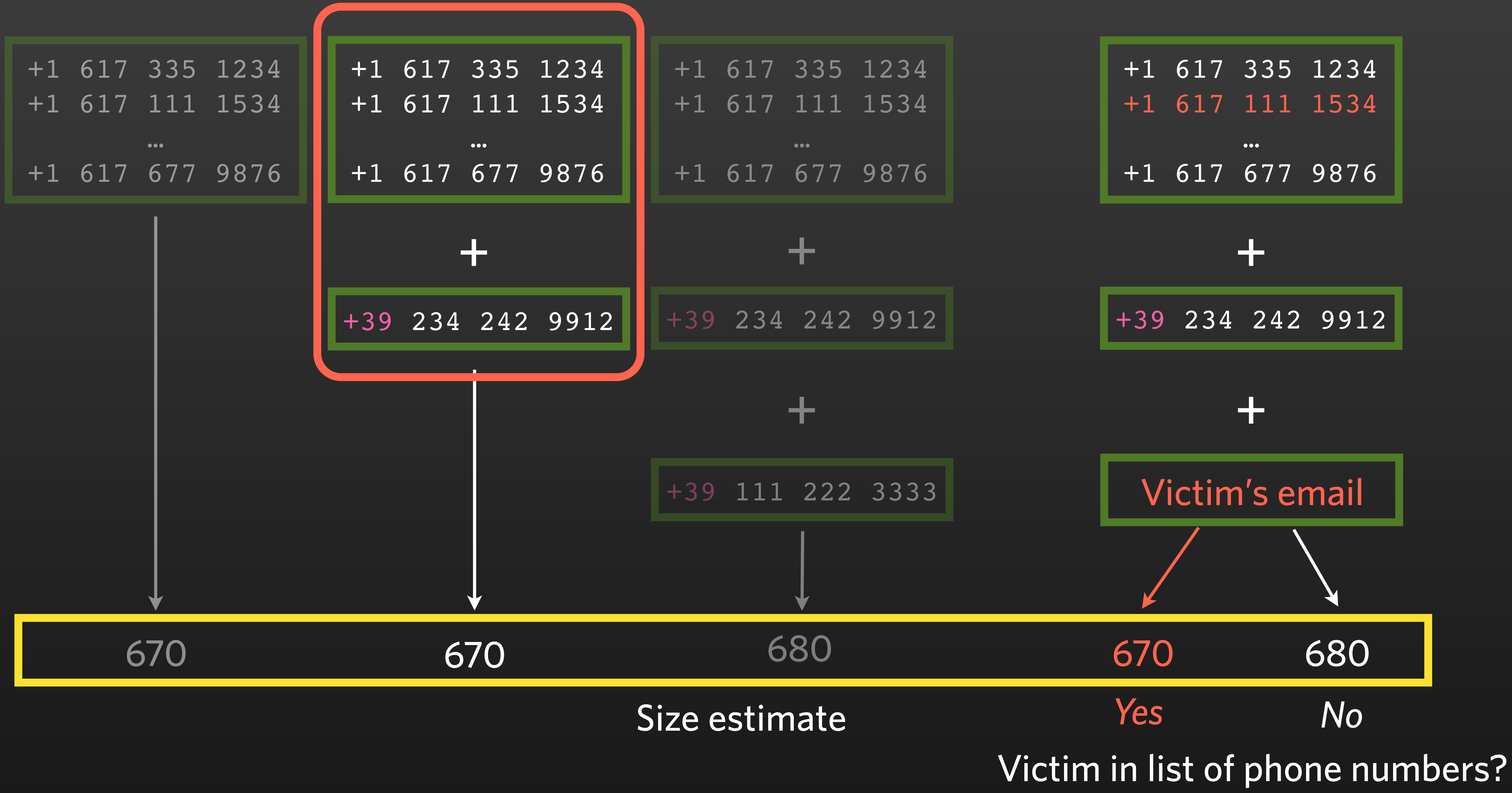
Victim's email

680

Size estimate

EXPLOITING THESE FEATURES

Goal: Given victim's email address, find if victim is in a given list of phone numbers



ATTACK: LEARNING USERS' PHONE NUMBERS

Can ask: Is Victim in Target List ?

ATTACK: LEARNING USERS′ PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

```
100-000-0000
100-000-0001
100-000-0002
     ...
199-999-9998
199-999-9999
```

ATTACK: LEARNING USERS' PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

```
100-000-0000
100-000-0001
100-000-0002
     ...
199-999-9998
199-999-9999
```

If ***No***: First digit is not 1

If ***Yes***: First digit is 1

ATTACK: LEARNING USERS' PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

100-000-0000
100-000-0001
100-000-0002
...
199-999-9998
199-999-9999

Is Victim in

200-000-0000
200-000-0001
200-000-0002
...
299-999-9998
299-999-9999

If ***No***: First digit is not 1

If ***Yes***: First digit is 1

ATTACK: LEARNING USERS' PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

```
100-000-0000
100-000-0001
100-000-0002
    . . .
199-999-9998
199-999-9999
```

Is Victim in

```
200-000-0000
200-000-0001
200-000-0002
    . . .
299-999-9998
299-999-9999
```

If ***No***: First digit is not 1

If ***Yes***: First digit is 1

If ***No***: First digit is not 2

If ***Yes***: First digit is 2

ATTACK: LEARNING USERS′ PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

```
100-000-0000
100-000-0001
100-000-0002
    ...
199-999-9998
199-999-9999
```

Is Victim in

```
200-000-0000
200-000-0001
200-000-0002
    ...
299-999-9998
299-999-9999
```

Is Victim in

```
010-000-0000
010-000-0001
010-000-0002
    ...
919-999-9998
919-999-9999
```

If ***No***: First digit is not 1

If ***Yes***: First digit is 1

If ***No***: First digit is not 2

If ***Yes***: First digit is 2

ATTACK: LEARNING USERS' PHONE NUMBERS

Can ask: Is Victim in Target List ?

Is Victim in

```
100-000-0000
100-000-0001
100-000-0002
     ...
199-999-9998
199-999-9999
```

If ***No***: First digit is not 1

If ***Yes***: First digit is 1

Is Victim in

```
200-000-0000
200-000-0001
200-000-0002
     ...
299-999-9998
299-999-9999
```

If ***No***: First digit is not 2

If ***Yes***: First digit is 2

Is Victim in

```
010-000-0000
010-000-0001
010-000-0002
     ...
919-999-9998
919-999-9999
```

If ***No***: Second digit is not 1

If ***Yes***: Second digit is 1

ROBUST DEFENSES: AN OPEN PROBLEM

Differential privacy

Poor fit to problem: hard/impossible to limit queries

Users can easily make additional accounts, use compromised accounts, etc

Size estimate obfuscation

Remove all size estimates

Coarse-grained size estimates (e.g., rounding)

Adding noise to size estimates
(e.g., to provide differential privacy)

Barrier to attacker

Require approval process for advertisers

Financial disincentives (e.g., pay per query)

Rate-limiting queries

Anomaly detection

Above defenses can be circumvented or lead to high loss of utility !

Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations

[NDSS'18]

EXPLANATIONS

Coming challenge: explain *why* a system made a particular decision

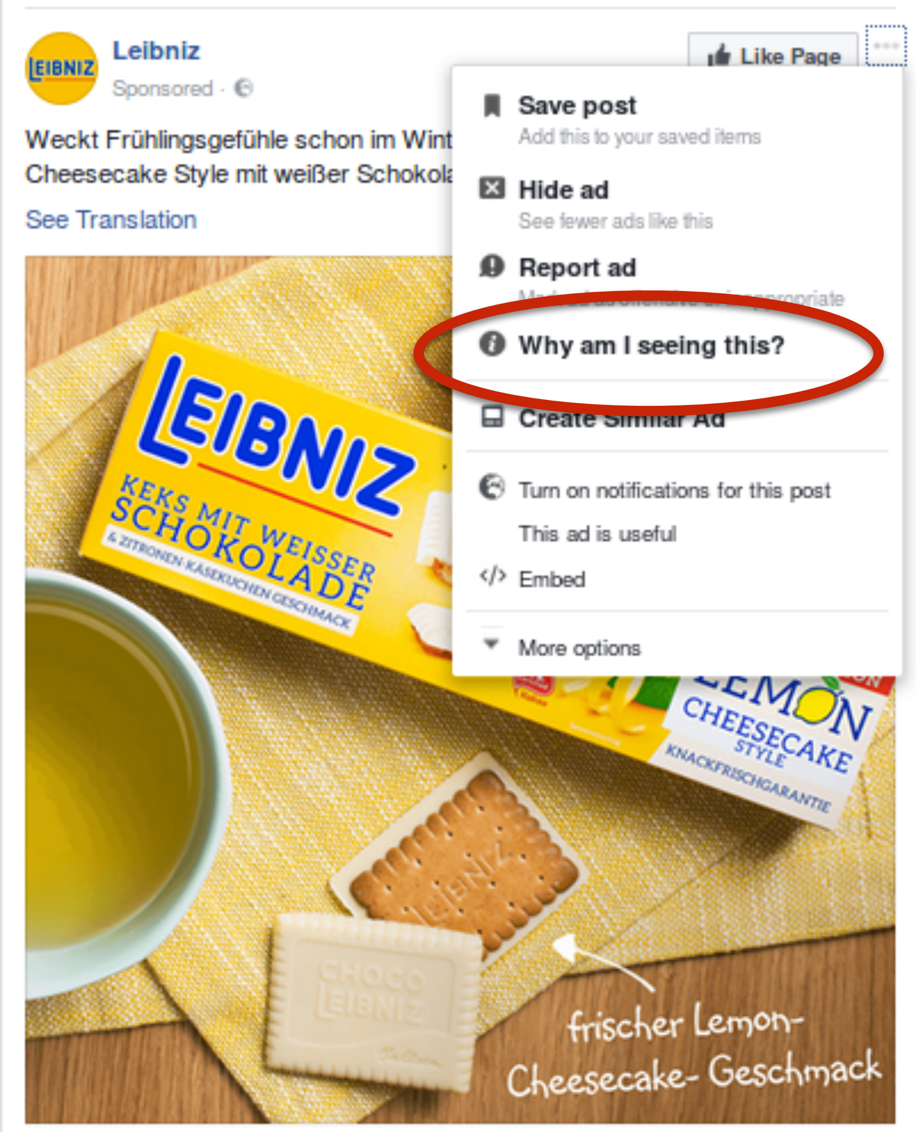
- Strong connections to data provenance

GDPR, French *Loi Numérique* may provide a “right to explanation”

Unclear even what makes a good explanation

- Who is the audience?
- What is the purpose?
- What are the privacy/security concerns?

Facebook already offers explanations for ads!



Leibniz

Sponsored ·

Like Page

Weckt Frühlingsgefühle schon im Wint
Cheesecake Style mit weißer Schokol

See Translation

Save post
Add this to your saved items

Hide ad
See fewer ads like this

Report ad
Mark as offensive or inappropriate

Why am I seeing this?

Create Similar Ad

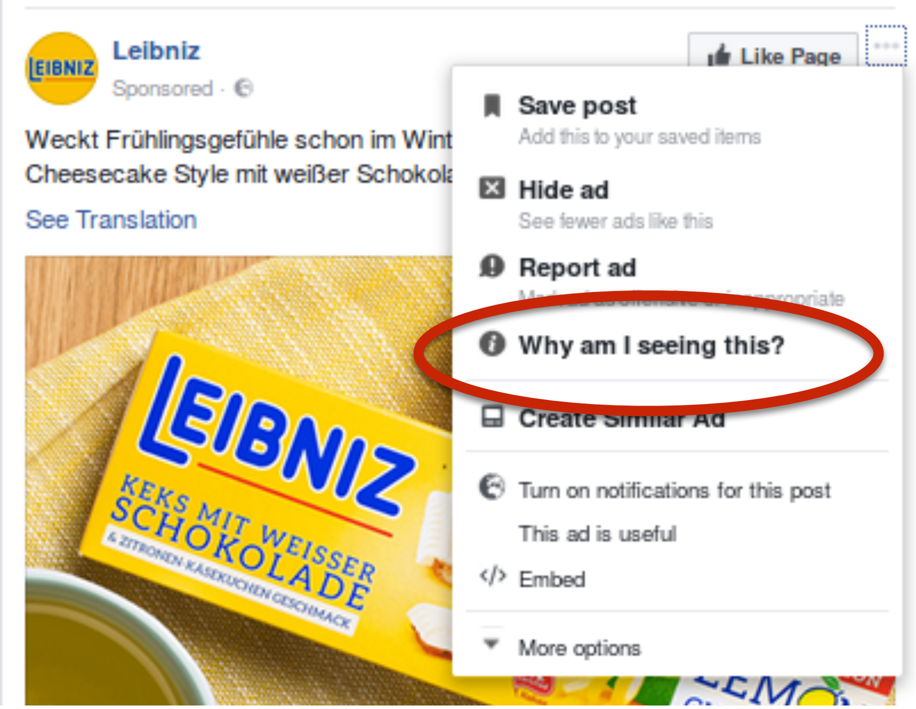
Turn on notifications for this post

This ad is useful

</> Embed

More options

LEIBNIZ
LEIBNIZ
KEKS MIT WEISSER SCHOKOLADE
& ZITRONEN-KÄSEKUCHEN GESCHMACK
LEMON
CHEESECAKE STYLE
KNACKFRISCHGARANTIE
LEIBNIZ
CHOCO LEIBNIZ
frischer Lemon-Cheesecake-Geschmack



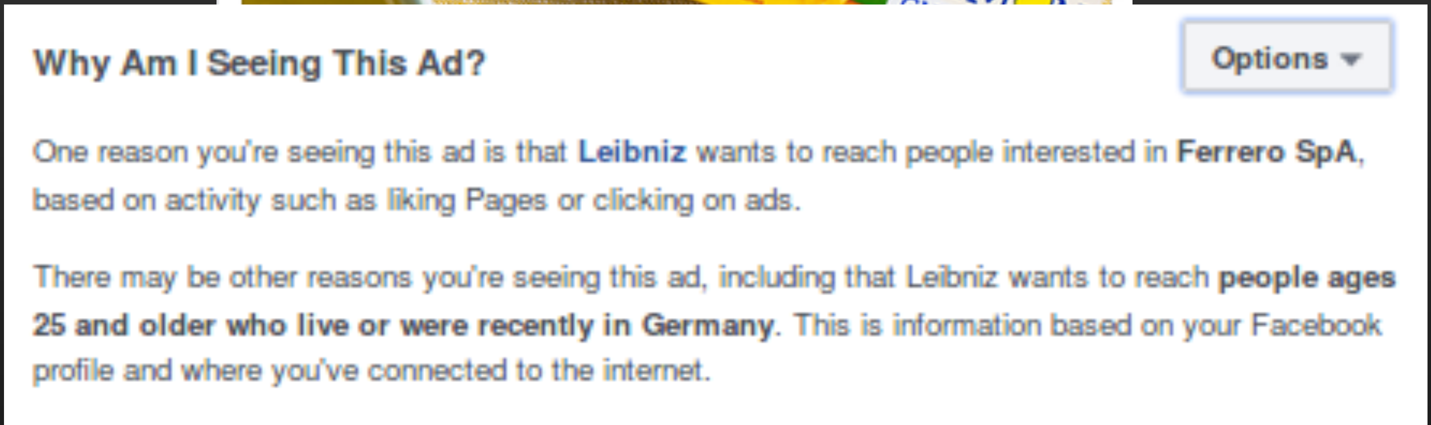
Leibniz

Sponsored ·

Weckt Frühlingsgefühle schon im Wint
Cheesecake Style mit weißer Schokola

See Translation

Like Page
Save post
Add this to your saved items
Hide ad
See fewer ads like this
Report ad
Mark ad as offensive or inappropriate
Why am I seeing this?
Create Similar Ad
Turn on notifications for this post
This ad is useful
Embed
More options
LEIBNIZ
KEKS MIT WEISSER SCHOKOLADE
& ZITRONEN-KÄSEKUCHEN GESCHMACK



Why Am I Seeing This Ad?

One reason you're seeing this ad is that **Leibniz** wants to reach people interested in **Ferrero SpA**, based on activity such as liking Pages or clicking on ads.

There may be other reasons you're seeing this ad, including that Leibniz wants to reach **people ages 25 and older who live or were recently in Germany**. This is information based on your Facebook profile and where you've connected to the internet.

EXPLANATIONS

Coming challenge: explain *why* a system made a particular decision

- Strong connections to data provenance

GDPR, French *Loi Numérique* may provide a “right to explanation”

Unclear even what makes a good explanation

- Who is the audience?
- What is the purpose?
- What are the privacy/security concerns?

Facebook already offers explanations for ads!

- Goal: Understand how explanations constructed, their properties
  - Correct? Complete? Misleading? Consistent?

METHODOLOGY AND RESULTS

Built browser extension to collect FB ads, explanations

- 35 users for 5 months
- 26K unique ads and corresponding explanations

Also ran controlled experiments with ads

- Targeted our 35 users with 96 different targeting parameters

Found that Facebook's explanations:

- *Personalized* — differ if users have different attributes
- *Incomplete* — have at most 1 targeting attribute, none from data brokers
- *Misleading* — use "may be other reasons" when there are not

USEFUL EXPLANATIONS: AN OPEN PROBLEM

Complete explanation for “why did I see this” would include:

- User’s attributes, all other users’ attributes
- Bids from all advertisers
- History of previous ad campaigns (calculating CTR)
- Implementation details of Facebook’s auction mechanism
- …

Open problem: constructing explanations for particular purposes

- What would auditors/regulators need?
- How to trade off complexity/utility for users?
- How to protect privacy?

QUESTIONS?