Privacy and PII in relation to the web

Tomas Persson 2021-11-05

Digitalist

https://digitalist.se/matomo

matomo camp

Who I am

https://github.com/tomper00

https://www.linkedin.com/in/tomper00/

Who we are

Digitalist

a digital product agency

that loves Open Source

Because we know that good ideas multiply when shared

EU

USA
(or other state)

Sweden

Citizen

stockholm.se

agreement

Stockholm.se
service vendors

Hosting

development

Laws

GDPR

OSL
"Swedish publicity and secrecy law"

Office 365

agreement

Laws

FISA 702

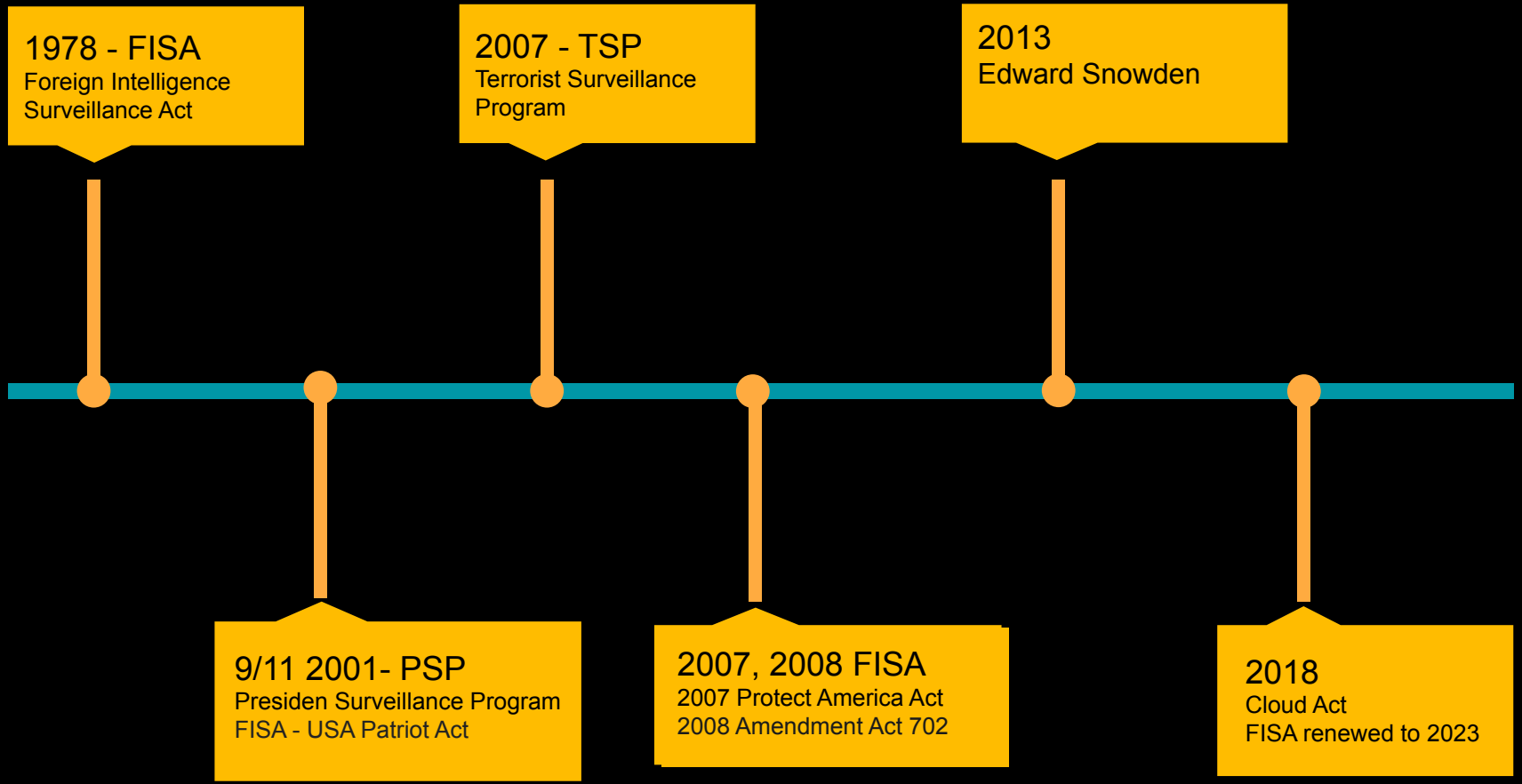
Microsoft

What are the actual problems with Google analytics and other services?

- Data leakage to foreign power
- Collection of personal data without consent
- We are selling data about our visitors

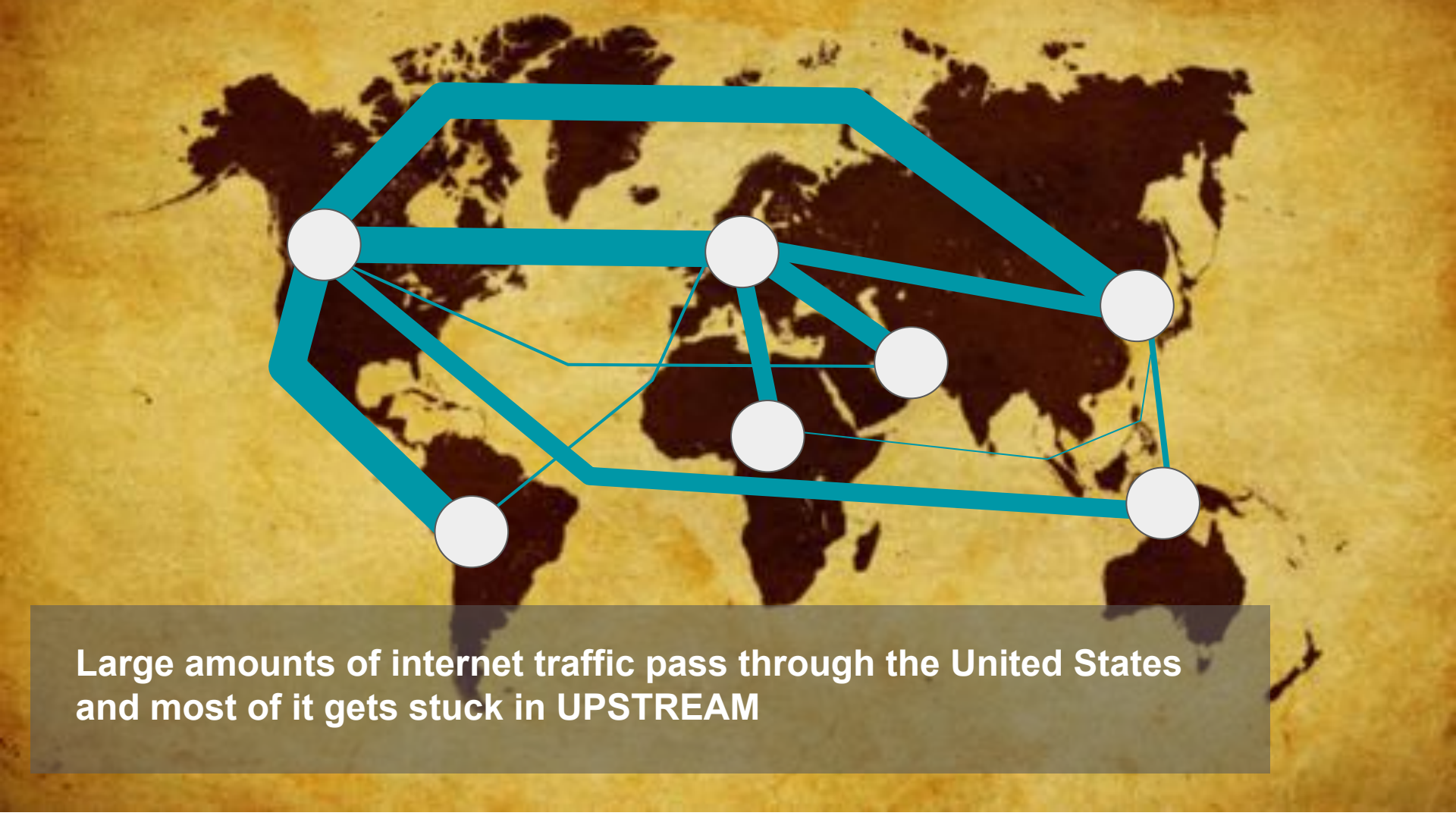
Digitalist


Data leakage to foreign power
Edward Joseph Snowden



Källa: https://www.eff.org/


Upstream - eavesdropping.
PRISM - data collection from companies.
TREASUREMAP - real-time info from phones etc ..
William "Bill" Binney, former NSA technical director
https://youtu.be/uYg_0lmrnr4



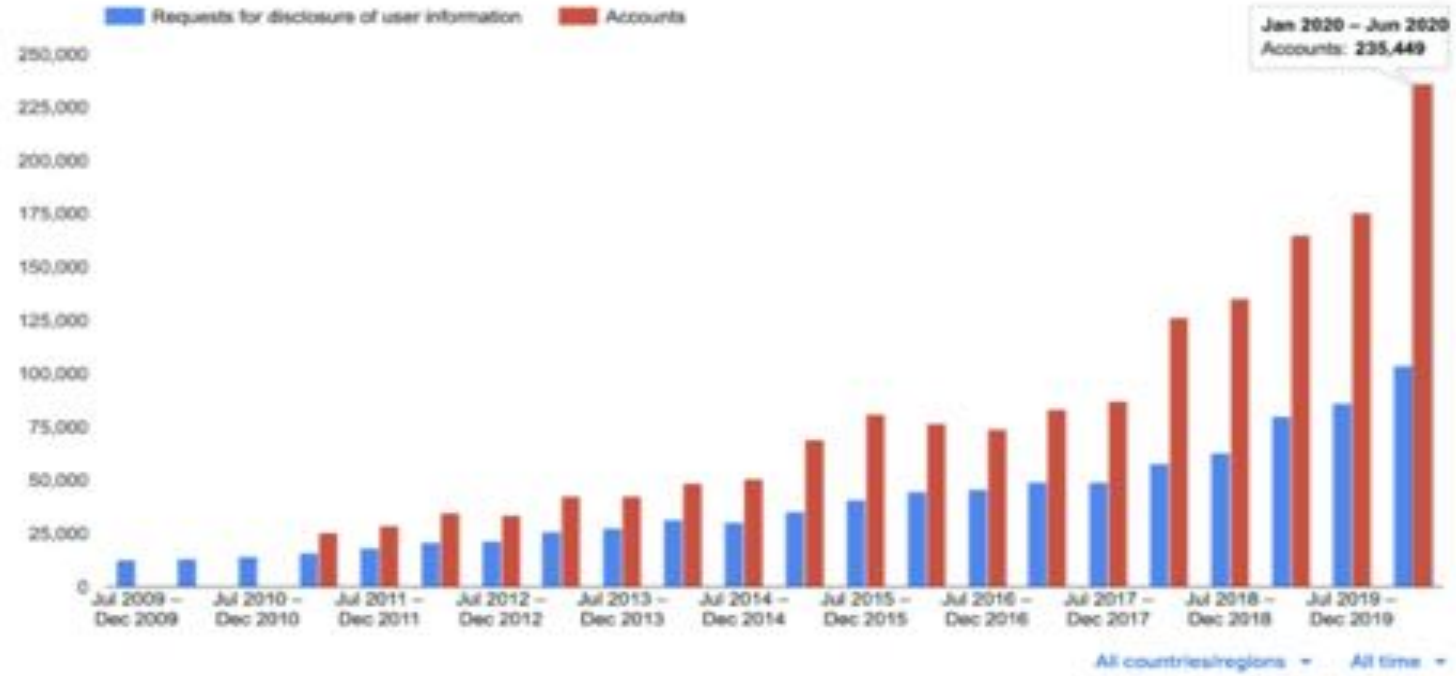
Large amounts of internet traffic pass through the United States and most of it gets stuck in UPSTREAM

Is this really a problem in a country like Sweden?

Privacy Reports

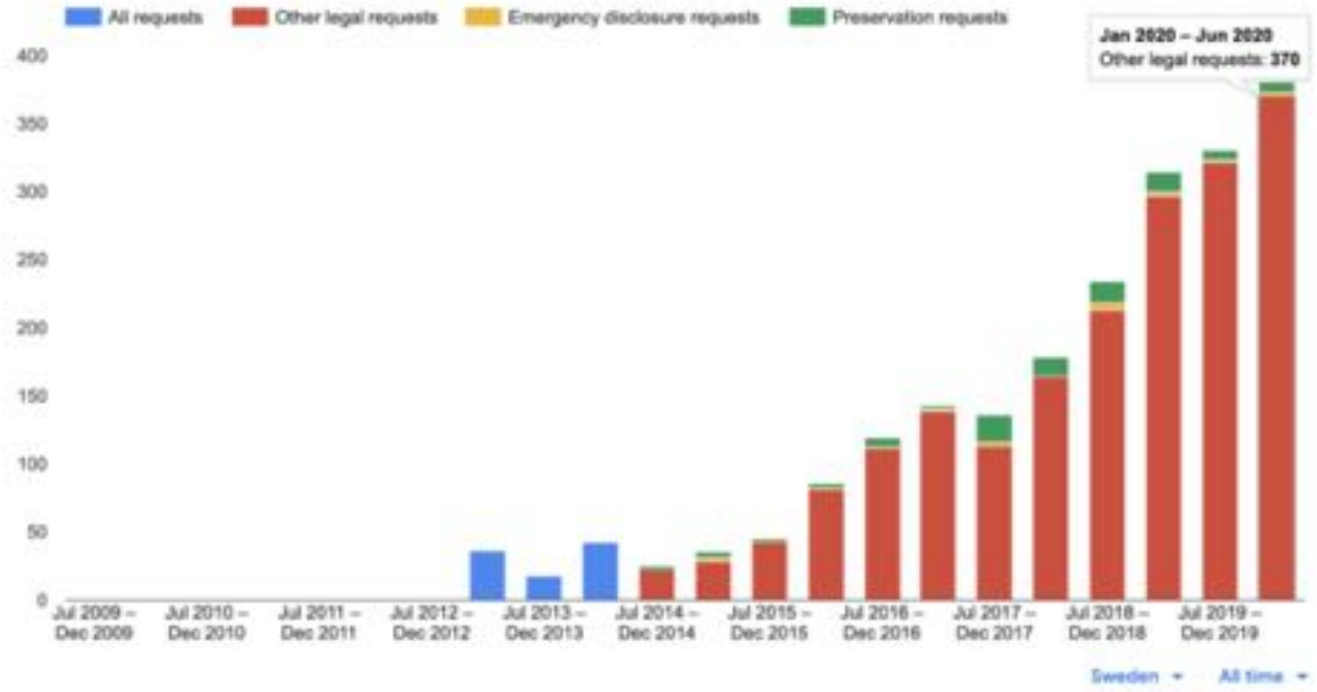
We publish this information to shine light on the impact government actions have on our users and the free flow of information online.

Google - Data om 235.449 användare (6 månader)



https://transparencyreport.google.com/user-data/overview?hl=en

370 in Sweden in the same period



https://transparencyreport.google.com/user-data/overview?hl=en&user_requests_report_period=series:requests,accounts;authority:SE;time:&lu=user_requests_report_period

Summary last half of 2019 in Sweden

Google - **321** - requests <--- Thus increased to 370 first half year (2021)

Microsoft - **396** - requests

Apple - **92** - units

Facebook - **548** - accounts

≈ 8 cases / day in Sweden

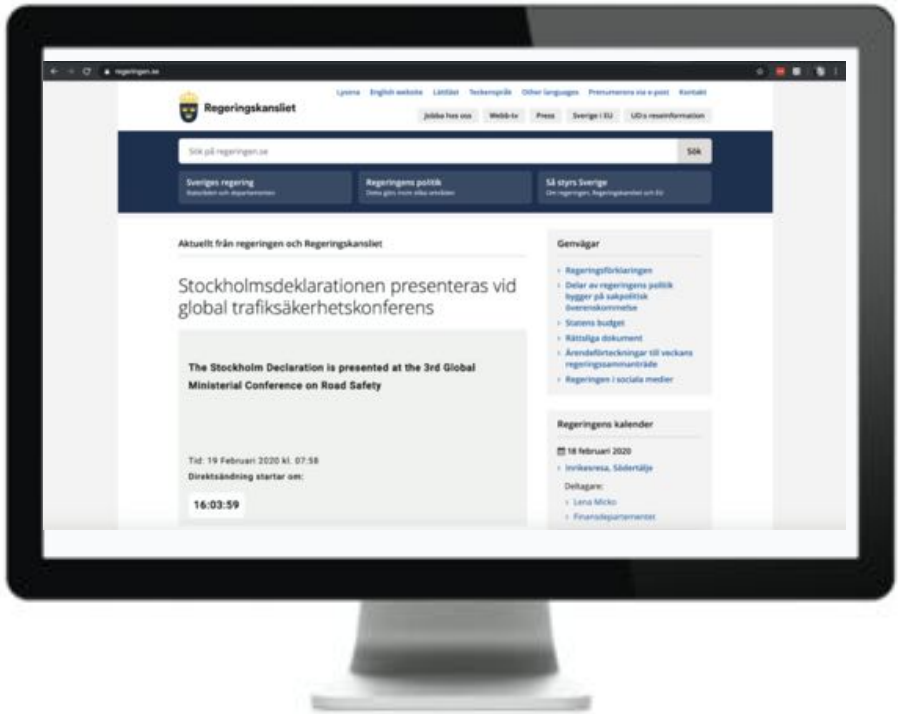
https://www.microsoft.com/en-us/corporate-responsibility/law-enforcement-requests-report
https://www.apple.com/legal/transparency/
https://govtrequests.facebook.com/
https://transparencyreport.google.com/user-data/overview?hl=en

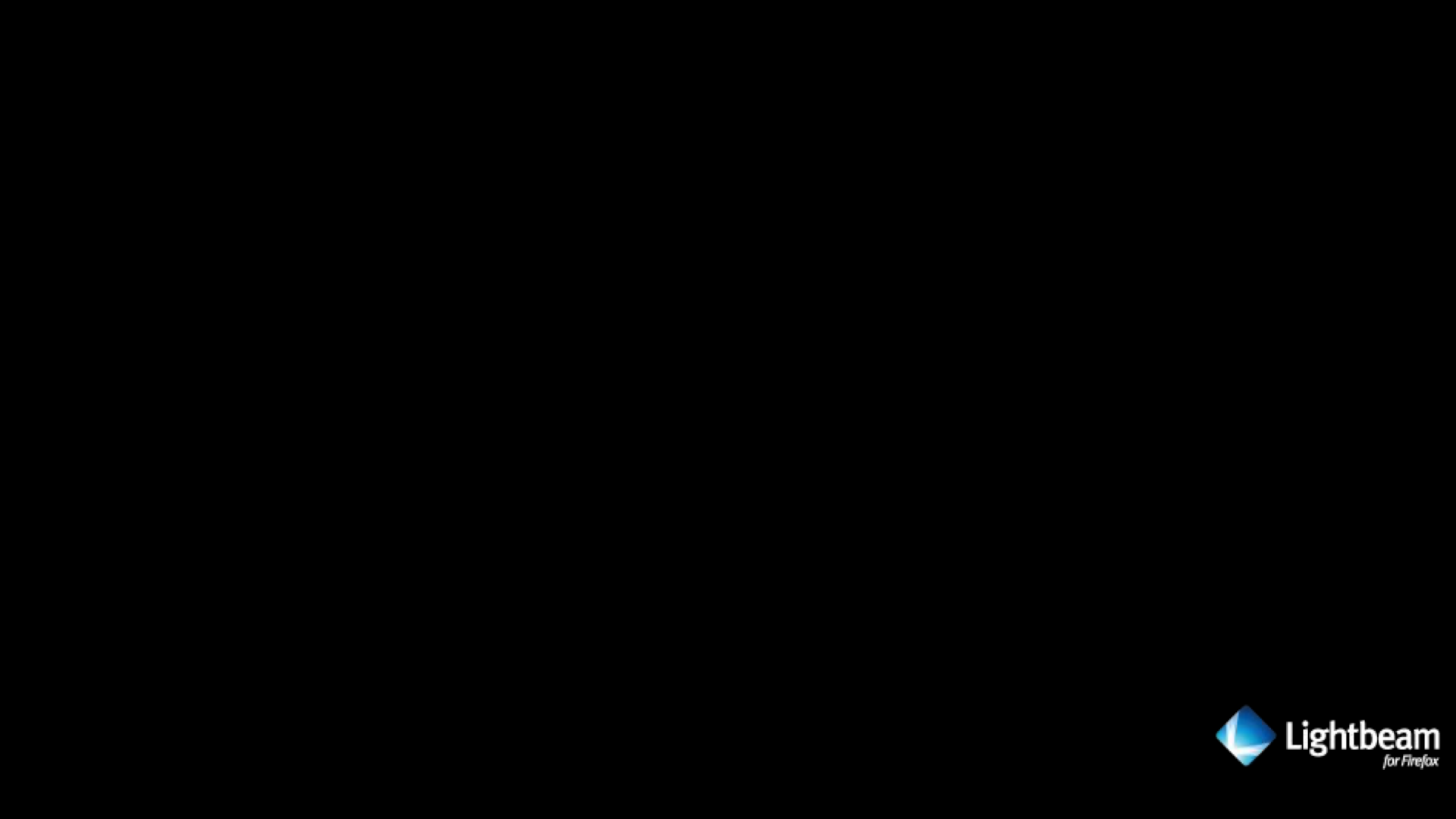
What does the public sector in Europe need to do?

- Secure the use of services (open technologies)
- Secure where the services run and who has access to them (secure cloud services)
- Secure traffic between residents and authorities (so that it never leaves Europe)

Leakage of personal data online web

What happens on the web when we visit 5 sites?




Lightbeam
for Firefox

Remember - Google sells info about our visitors

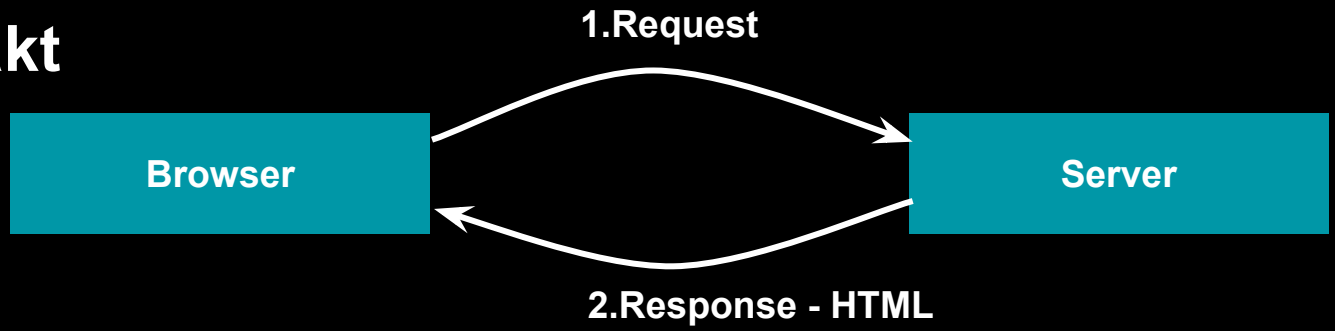
83%

But Google Analytics is only a small part of the bigger problem

regeringen.se/kontakt/

A technical deep dive at the Swedish Gov

/kontakt



33 request to 6 domains

kontakt/	200	www.regeringen.se	document
modernizr.custom.min.30492.js	200	www.regeringen.se	script
bundle?v=p8MLSOU6gyQRqs8xcM3Li7kuA0_vXON-nMGlvNY0_p01	200	www.regeringen.se	stylesheet
GoBrain.min.js	200	play2.qbrick.com	script
logo-sv.png	200	www.regeringen.se	png
find.js	200	dl.episerver.net	script
js?v=LWrcM8g6WfrwT3dPOO3DTI0Pssi88krr8O_hPGyHBx1g1	200	www.regeringen.se	script
ReadSpeaker.js?pids=embhl	200	www.regeringen.se	script
ajax-loader.gif	200	www.regeringen.se	gif
gtm.js?id=GTM-P4L58V	200	www.googletagmanager.com	script
opensans-regular-webfont.woff2	200	www.regeringen.se	font
opensans-semibold-webfont.woff2	200	www.regeringen.se	font
opensans-light-webfont.woff2	200	www.regeringen.se	font

Request

Request URL: https://dl.episerver.net/13.2.5/epi-util/find.js

Request Method: GET

Status Code: 200

Remote Address: 104.18.18.118:443

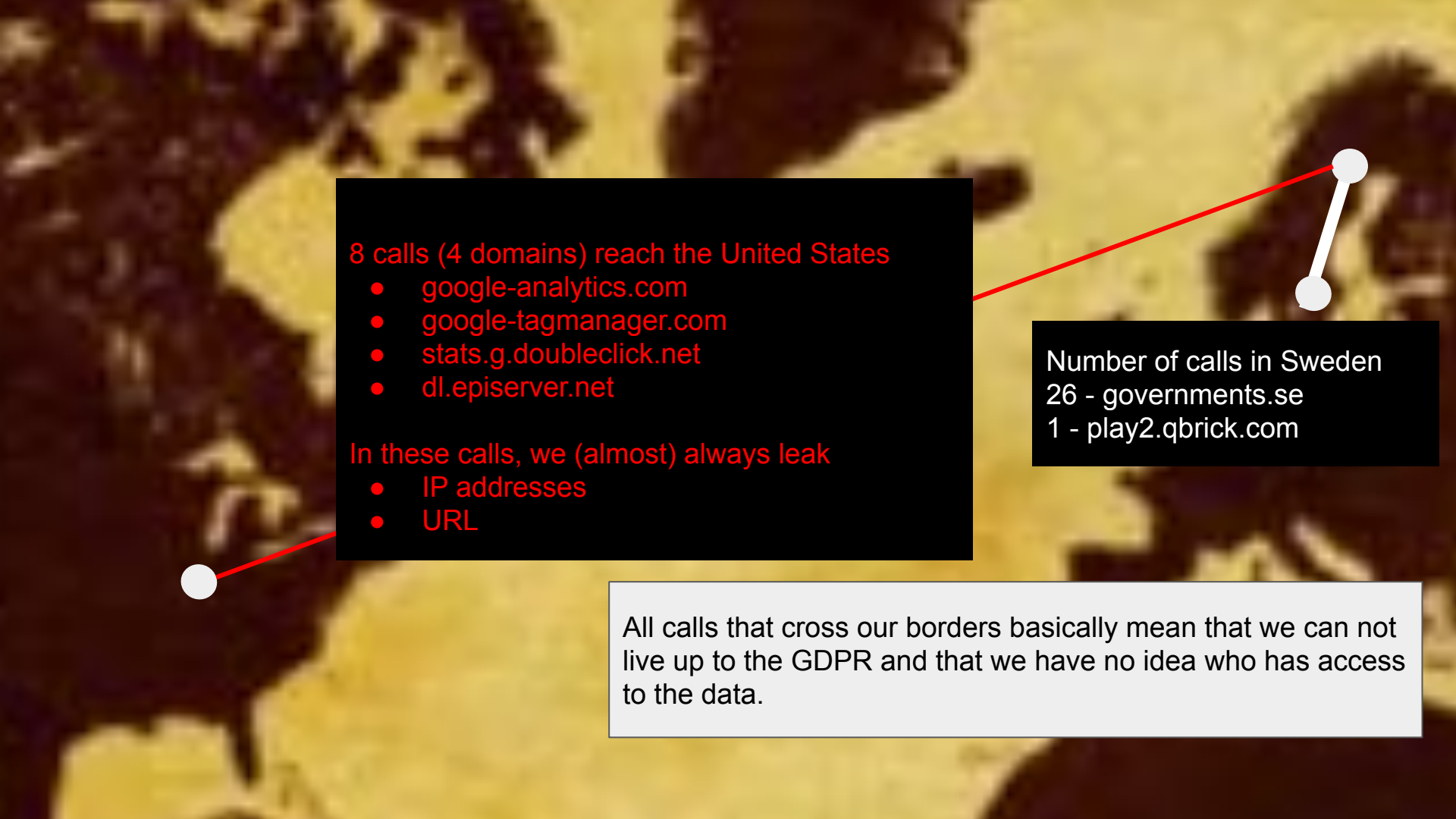
Referrer Policy: no-referrer-when-downgrade

Request headers

:authority: dl.episerver.net
:method: GET
:path: /13.2.5/epi-util/find.js
:scheme: https
accept: */*
accept-encoding: gzip, deflate, br
accept-language: sv-SE,sv;q=0.9,en-US;q=0.8,en;q=0.7
cache-control: no-cache
cookie: __cfduid=d9688888259e51a1e4d0cf79cc34596c6157
pragma: no-cache
referer: https://www.regeringen.se/kontakt/
sec-fetch-mode: no-cors
sec-fetch-site: cross-site
user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_

kontakt/	www.regeringen.se
modernizr.custom.min.30492.js	www.regeringen.se
bundle?v=p8MLSOU6gyQRqi8xcM3Li7kuA0_vXON-nMGlvNY0_p01	www.regeringen.se
GoBrain.min.js	play2.qbrick.com
logo-sv.png	www.regeringen.se
find.js	dl.episerver.net
js?v=LWrcM8g6WfrwT3dPOO3DTl0Pssi88km8O_hPGyHBx1g1	www.regeringen.se
ReadSpeaker.js?pids=embhl	www.regeringen.se
ajax-loader.gif	
gtm.js?id=GTM-P4L58V	
opensans-regular-webfont.woff2	
opensans-semibold-webfont.woff2	
opensans-light-webfont.woff2	

The information shared in headers means that all servers we call can do the same type of analysis that we have in Google Analytics by default. (Pageviews with segmentation on technologies, IP addresses etc.)


8 calls (4 domains) reach the United States
google-analytics.com
google-tagmanager.com
stats.g.doubleclick.net
dl.episerver.net
In these calls, we (almost) always leak
IP addresses
URL
Number of calls in Sweden
26 - governments.se
1 - play2.qbrick.com
All calls that cross our borders basically mean that we can not live up to the GDPR and that we have no idea who has access to the data.

A lot of other services actually have the same issues.

Other commonly used scripts with the same problem:

- Readspeaker
- SiteImprove statistics
- Apsis
- EpiServer find
- NewRelic
- etc

PII

(Personally identifiable information)

- IP addresses
- Email addresses
- ZIP code
- Social security number
- Credit card number
- Name
- Phone number
- etc.

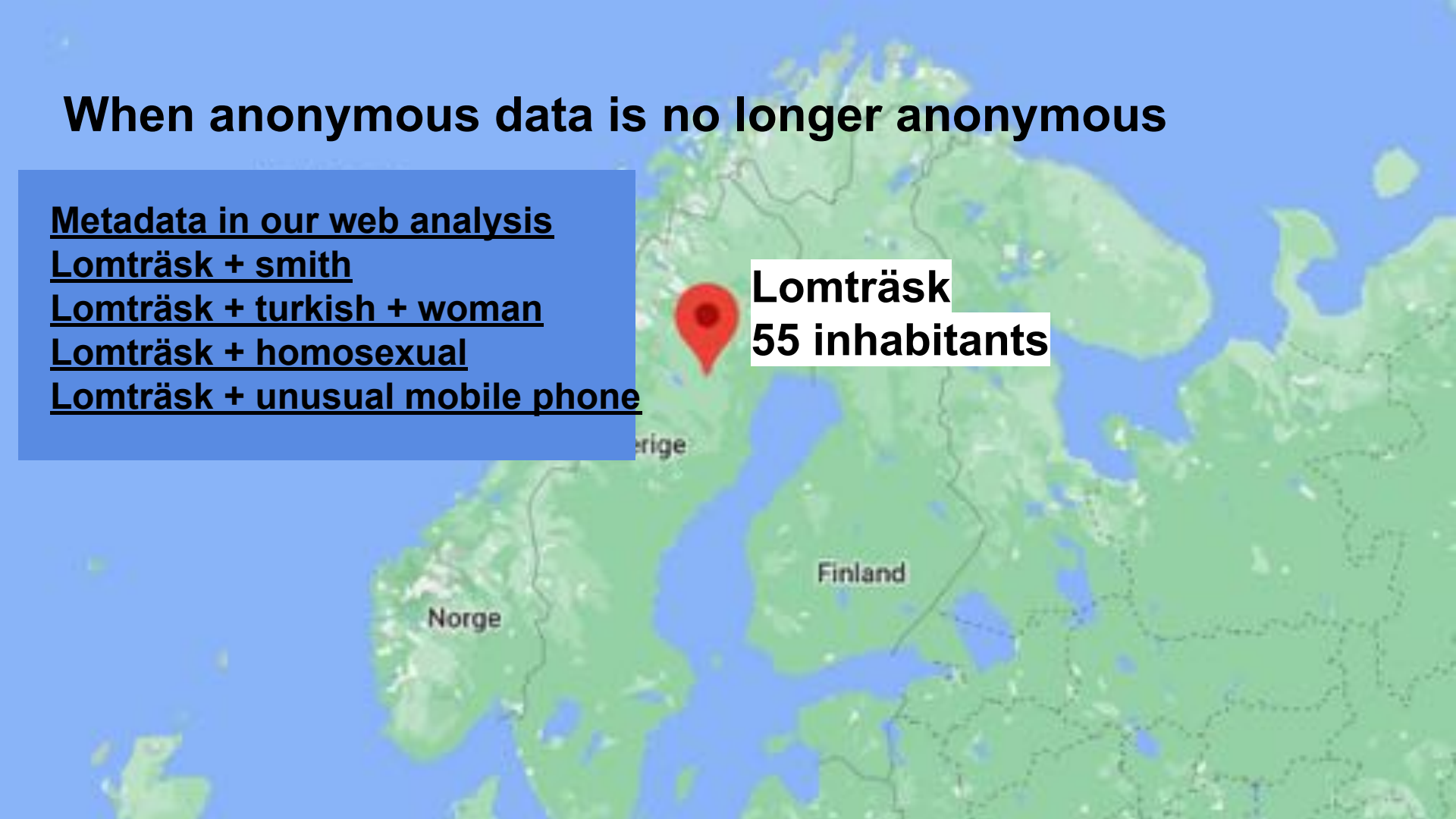
Personal data

Personal data is also

Data that in combination with other data can identify a person

When anonymous data is no longer anonymous

Metadata in our web analysis
Lomträsk + smith
Lomträsk + turkish + woman
Lomträsk + homosexual
Lomträsk + unusual mobile phone



Lets privacy test your site

https://github.com/tomper00/privacy-test-your-site


What do we do when
the accident occurs?

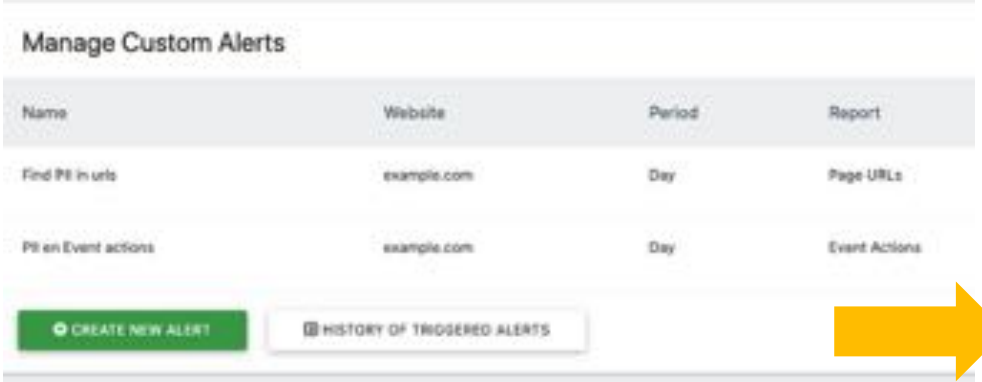
There must always be routines in place that assume that we will find PII.


PII incident
Example of action Plan
1. Assess severity
2. Prevent further collection
3. Decide on any report to your data protection authority (within 72 hours)
4. Clear data
5. Document incidents
6. Inform victims?
7. Retrospective

Deleting data in Matomo

1. Find the data
2. Find the visitor id´s
3. Delete the visits
4. Reprocess the visitor log so that aggregated data is updated

Simple example of a simple PII monitoring with the Alert Plugin



Example regex:
Finding email-addresses:
^[^\s@]+@[^\s@]+\.[^\s@]

Finding dates
$|\d\d\d\d-\d\d\-\d\d-\d\d

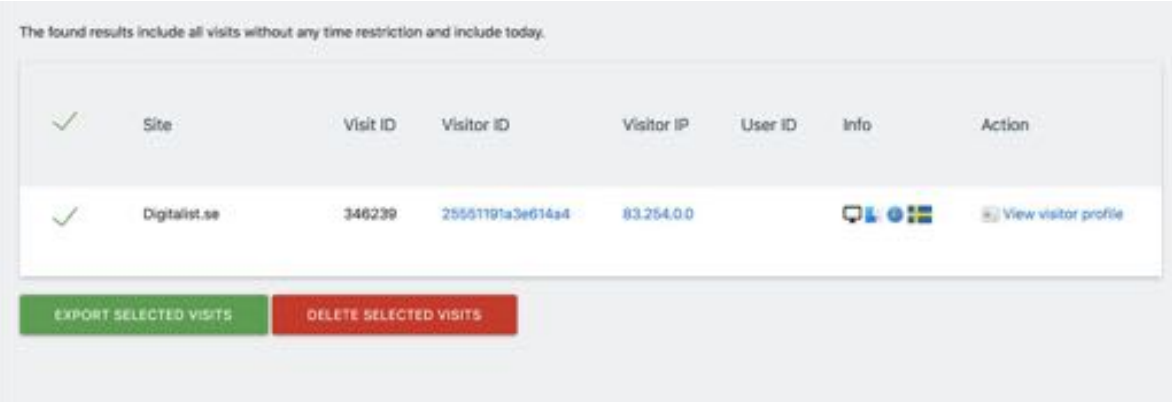
https://plugins.matomo.org/CustomAlerts

Combine multiple checks with |
^[^\s@]+@[^\s@]+\.[^\s@]|$|\d\d\d\d-\d\d\-\d\d-\d\d

Drawback:
Alarms comes in (the next day)
Hard to detect things like a name in negex

Deleting data with GDPR Tools or the API

Search for visitors (but very limited) since we will not be able to find users matching an event name containing tomas@digitalist.se



Deleting a lot of data

https://matomo-server/index.php?module=API&method=PrivacyManager.deleteDataSubjects&visits[0][idsite]=1&visits[0][idvisit]=**12345**&token_auth=API_TOKEN

Future ideas - create a PII monitoring plugin

Set up rules to find PII (on a wider scale) on Events, Pageviews, referrer urls etc

Set up alarms

Make it possible to add exceptions (when we find thinks that are ok)

Make it easy to:

- Delete data (both in visitor logs and aggregates)
- Anonymize data (both in visitor logs and aggregates)

6 steps to secure your web analysis

1. Collect consents and inform visitors
2. Check your data collection (avoid collection PII) and set up Matomo properly
3. Manage the quality on your applications and websites
4. Secure your technical infrastructure for your system and control who has access
5. Limit time for how long data is stored
6. Processes in place to handle incidents and to monitor your data

Questions