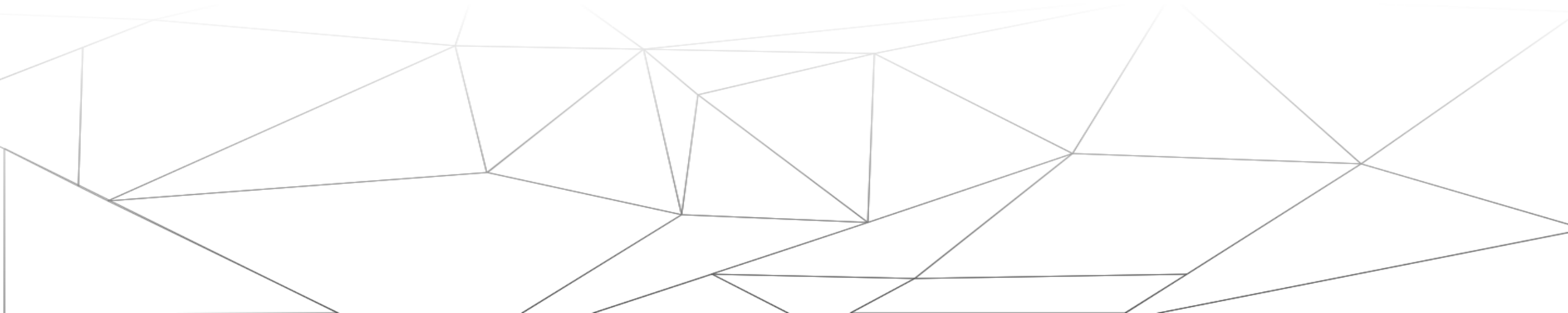


Dark Data

Svea Eckert – Andreas Dewes



Who we are

Svea Eckert

Journalist NDR/ ARD



@sveckert

Andreas Dewes
(data) scientist



@japh44

Why we are here



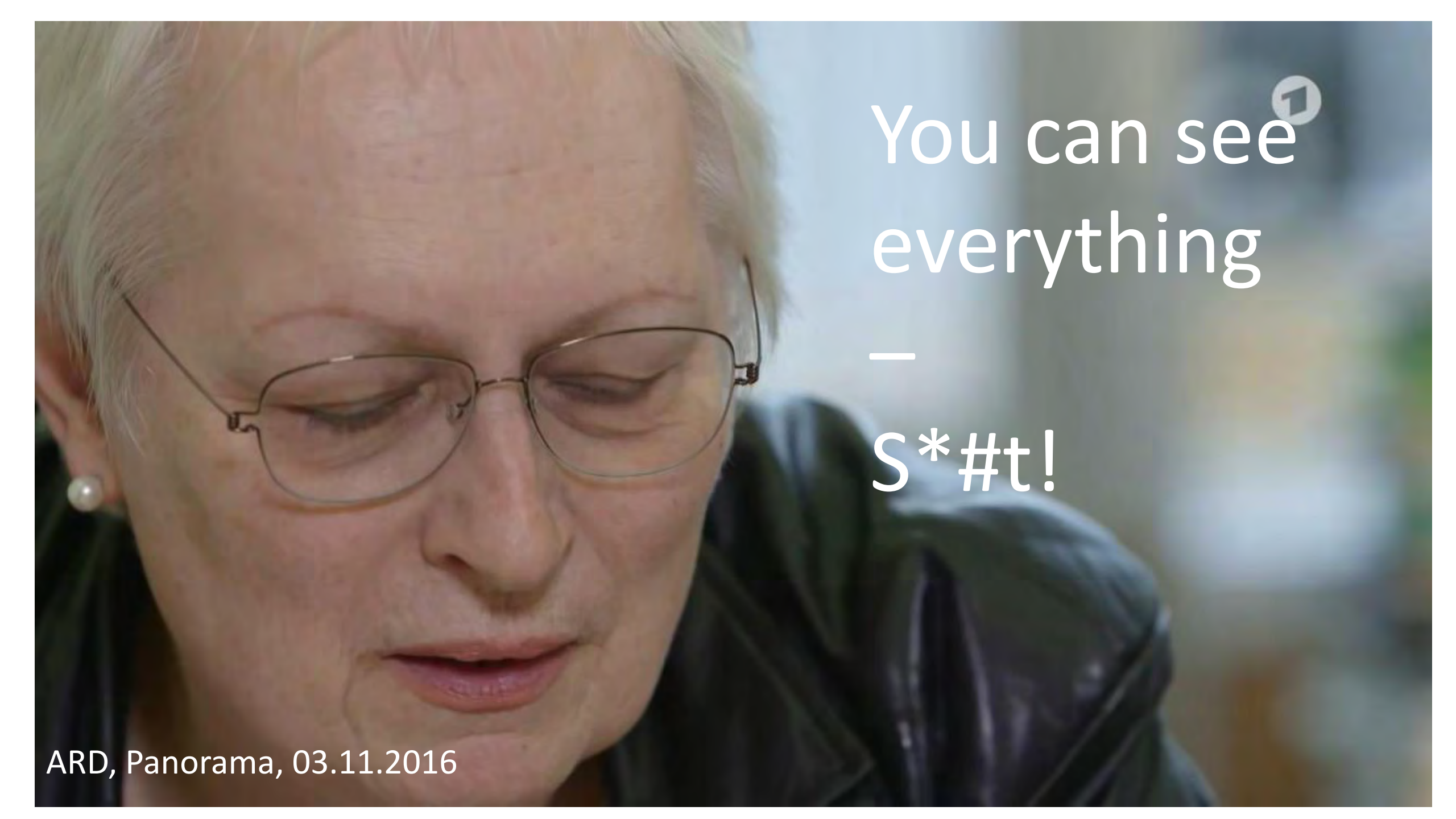
*“US Senate voted to eliminate **broadband privacy rules** that would have required ISPs to get consumers' explicit consent before selling or sharing **Web browsing data (...)**”*

3/23/2017

<https://arstechnica.com>

What does that mean





You can see¹
everything

—

S*#t!

01.08.16 06:11:48 https://banking.postbank.de/rai/crypt/kCvFlnxhWfHAIJhNSUZn9QVwsNnwtrB9mlJHL6OqFglQ9L_VVPY19RRP33jXOOcftP3aNTLgHu2s9xAQLOYFm9fPRthhSM

01.08.16 06:11:59 https://banking.postbank.de/rai/crypt/kCvFlnxhWfHAIJhNSUZn9QVwsNnwtrB9mlJHL6OqFglQ9L_VVPY19RRP33jXOOcfKHbzSQMjLWgdC2tnqsW0v19fw13mE

01.08.16 06:12:06 | https://banking.postbank.de/rai/crypt/kCvFlnxhWfHAIJhNSUZn9QVwsNnwtrB9mlJHL6OqFglQ9L_VVPY19RRP33jXOOcftP3aNTLgHu2s9xAQLOYFm9fPRthhSM

01.08.16 06:12:45 <https://banking.postbank.de/rai/logout>

01.08.16 06:13:10 <https://www.schwaebisch-hall.de/>

01.08.16 06:13:16 | <https://www.schwaebisch-hall.de/mein-konto.html#contracts>

01.08.16 06:13:24 <http://www.welt.de/>

01.08.16 06:13:43 <http://www.welt.de/>

01.08.16 06:23:55 <http://www.sueddeutsche.de/>

01.08.16 06:24:42 <https://outlook.live.com/owa/?path=/mail/inbox/rp>

01.08.16 06:25:47 <http://www.spiegel.de/auto/fahrkultur/gm-futurliner-der-science-fiction-bus-von-general-motors-a-1102970.html>

01.08.16 06:26:27 <http://www.spiegel.de/>

01.08.16 06:27:38 <http://www.spiegel.de/politik/ausland/news-des-tages-sigmar-gabriel-horst-seehofer-erdogan-demo-a-1105493.html>

01.08.16 06:27:47 <http://m.welt.de/wirtschaft/article157421718/Der-gefaehrliche-Geiz-im-Containergeschaeft.html>

01.08.16 06:27:53 <http://m.welt.de/wirtschaft/article157421718/Der-gefaehrliche-Geiz-im-Containergeschaeft.html>

02.08.16 16:47:34 <http://www.welt.de/wirtschaft/article157421718/Der-gefaehrliche-Geiz-im-Containergeschaeft.html>

02.08.16 16:47:46 <http://www.welt.de/wirtschaft/article157421718/Der-gefaehrliche-Geiz-im-Containergeschaeft.html?config=print>

02.08.16 16:47:50 <http://www.welt.de/wirtschaft/article157421718/Der-gefaehrliche-Geiz-im-Containergeschaeft.html>

02.08.16 16:48:32 <https://www.freitag.de/autoren/felix-werdermann/das-autonome-auto>

02.08.16 16:49:49

02.08.16 16:53:36 <https://banking.postbank.de/rai/login>

02.08.16 17:00:09 <https://banking.postbank.de/rai/crypt/25KzWw0hDqwcUJC5ZH96GpdbmQV6DusTifPKNOk4CUQoj3sDon5x7hiWa3NOr6uIEUrQrHA3rC3rjyKReCIIOfnPEWvOLv>

02.08.16 17:00:22 <https://banking.postbank.de/rai/crypt/6BHPwWrPQKHP1iv7QC191A/6BH8b>

02.08.16 17:00:28 <https://banking.postbank.de/rai/logout>

02.08.16 17:00:45 <https://www.schwaebisch-hall.de/>

02.08.16 17:00:51 <https://www.schwaebisch-hall.de/mein-konto.html#contracts>

02.08.16 17:00:57 <https://www.schwaebisch-hall.de/hinweise/logout.html>

02.08.16 17:01:41 <http://www.spiegel.de/>

02.08.16 17:01:50 <http://www.spiegel.de/politik/deutschland/petra-hinz-warum-gibt-spd-abgeordnete-nach-dem-betrug-ihr-mandat-nicht-zurueck-a-1105823.html>

02.08.16 17:01:56 <http://www.spiegel.de/>

02.08.16 17:05:51 <http://www.faz.net/>

02.08.16 17:06:32 <http://www.faz.net/aktuell/wirtschaft/vw-abgasskandal/abgas-skandal-gruene-kritisieren-csu-setzt-euch-fuer-deutsche-vw-kunden-ein-14368638.html>

24.08.16 16:50:07 <http://www.spiegel.de/>
24.08.16 16:52:46 <http://www.spiegel.de/wissenschaft/natur/italien-warum-erdbeben-das-ganze-land-bedrohen-a-1109183.html>
24.08.16 16:52:52 <http://www.welt.de/>
24.08.16 16:55:56 <http://www.welt.de/wirtschaft/article157839434/Dieses-Auto-rettet-die-Post-vorm-staedtischen-Liefer-GAU.html>
24.08.16 16:56:47 <http://www.abendblatt.de/>
24.08.16 16:58:52 <https://www.elsteronline.de/eportal/Logout.tax>
24.08.16 16:59:18 <https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?action=/eop/auth/LoginElsterSmart&requestaction=KonfigurationsAssistent.tax&clientTechnolo>
24.08.16 17:21:54 https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?ElsterRequestKeys.KONFIGURATIONS_ASSISTENT_WORKFLOW=BXJgdmqSdWlBm7K&sr
24.08.16 17:22:09 https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?ElsterRequestKeys.KONFIGURATIONS_ASSISTENT_WORKFLOW=CJAli4UJwiThfjU&schritt
24.08.16 17:22:09 <https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?requestaction=KonfigurationsAssistent1JS.tax&action=/eop/auth/LoginSoft-PSE-JS>
24.08.16 17:22:33 https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?ElsterRequestKeys.KONFIGURATIONS_ASSISTENT_WORKFLOW=HovqLgpLUoB1jaa&sch
24.08.16 17:22:33 <https://www.elsteronline.de/hilfe/eop/public/help.html>
24.08.16 17:22:40 <https://www.sicherheitsstick.de/>
24.08.16 17:24:07 <https://www.sicherheitsstick.de/technik.html>
24.08.16 17:25:11 <https://www.sicherheitsstick.de/technik.html>
24.08.16 17:26:02 https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?ElsterRequestKeys.KONFIGURATIONS_ASSISTENT_WORKFLOW=HovqLgpLUoB1jaa&sch
24.08.16 17:26:32 https://www.elsteronline.de/eportal/KonfigurationsAssistent1JS.tax?ElsterRequestKeys.KONFIGURATIONS_ASSISTENT_WORKFLOW=HovqLgpLUoB1jaa&sch
24.08.16 17:26:50 <https://www.elsteronline.de/eportal/SwitchModule.tax?switchpage=/Postfach.tax&switchprefix=/all/postfach>
24.08.16 17:36:38 <https://www.elsteronline.de/eportal/all/postfach/AufgabenVerschluesseltLaden.tax?aufgabeld=30873099&org.apache.struts.taglib.html.TOKEN=c9b64c02-a062-4>
24.08.16 17:36:48 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/FormularEst.tax>
24.08.16 17:36:50 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/Hauptvordruck.tax>
24.08.16 17:36:57 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckStpflPerson.tax>
24.08.16 17:36:59 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckLebenspartner.tax>
24.08.16 17:37:28 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckVeranlagungsart.tax>
24.08.16 17:37:36 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckBankverbindung.tax>
24.08.16 17:37:39 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckEmpfangsbevollmaechtigter.tax>
24.08.16 17:37:48 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckGezahlteVersorgungsleistungen.tax>
24.08.16 17:37:52 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckAusglZahlgUnterhaltsleistgKirchensteuer.tax>
24.08.16 17:38:01 <https://www.elsteronline.de/eportal/eop/formular/est/y2015/HauptvordruckAusglZahlgUnterhaltsleistgKirchensteuer.tax>

18.08.16 18:39:53 <http://www.sueddeutsche.de/wirtschaft/autoindustrie-vw-will-mit-voller-haerte-gegen-streikende-zulieferer-vorgehen-1.3126280-2>

18.08.16 18:41:37 <http://www.handelsblatt.com/>

18.08.16 18:43:22 <http://www.handelsblatt.com/unternehmen/industrie/volkswagen-werke-vw-schickt-ueber-20-000-beschaefigte-in-kurzarbeit/14026328.html>

18.08.16 18:43:24 <http://teslamag.de/>

18.08.16 18:45:02 <http://teslamag.de/news/leitender-mitarbeiter-teslas-navigation-unternehmen-9308>

18.08.16 18:45:17 <http://teslamag.de/>

18.08.16 18:46:07 <http://teslamag.de/news/model-3-umgeklappte-rueckbank-wird-eine-ebene-ladeflaeche-bieten-9332>

18.08.16 18:46:14 <http://teslamag.de/news/tesla-wird-die-eigenen-stores-radikal-umgestalten-sagt-verkaufsleiter-9329>

18.08.16 18:46:56 https://www.tesla.com/de_DE/

18.08.16 18:47:18 https://www.tesla.com/de_DE/models/design

18.08.16 18:47:27 <https://r.duckduckgo.com//?kh=-1&uddq=http%3A%2F%2Fwww.sanego.de%2FMedikamente%2FTebonin%2F>

18.08.16 18:47:53 <http://www.sanego.de/Medikamente/Tebonin/>

18.08.16 18:48:43

18.08.16 18:48:43

18.08.16 18:49:28 <https://duckduckgo.com/?q=tebonin+wikipedia&atb=v19&ia=web>

18.08.16 18:49:36 <https://de.wikipedia.org/wiki/Bilobalid>

18.08.16 18:49:53 https://de.wikipedia.org/wiki/Dr._Willmar_Schwabe

18.08.16 18:50:10

18.08.16 18:52:38 <http://www.weser-kurier.de/>

19.08.16 07:18:41 <http://www.weser-kurier.de/region/osterholzer-kreisblatt.html>

19.08.16 07:19:30

19.08.16 07:19:52

19.08.16 07:37:29 <https://duckduckgo.com/?q=b74-g10&atb=v19&ia=web>

19.08.16 07:38:23 <https://r.duckduckgo.com//?kh=-1&uddg=http%3A%2F%2Fwww.bvwp-projekte.de%2Fstrasse%2FB74-G10-NI%2FB74-G10-NI.html>

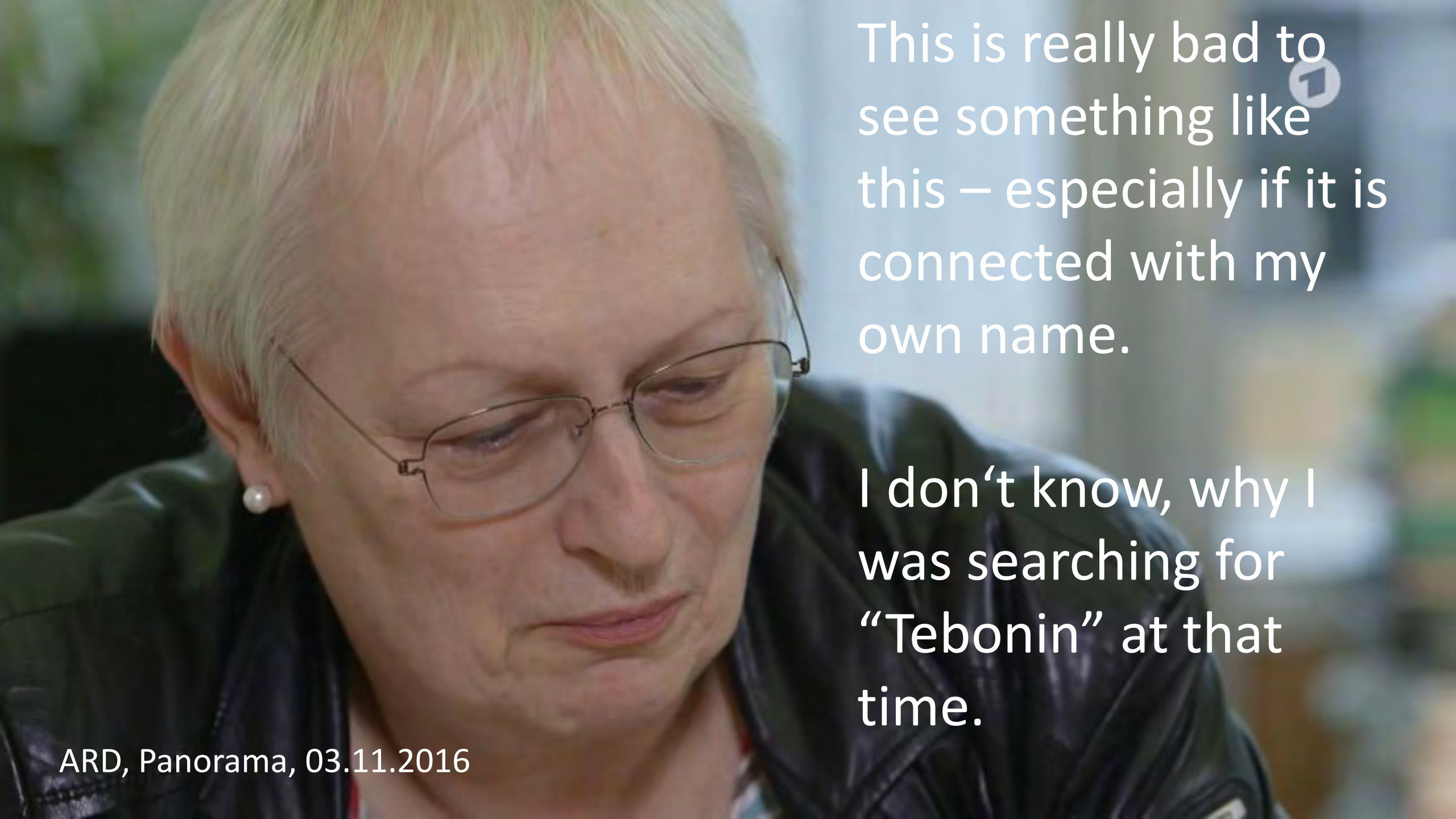
19.08.16 07:38:27 <http://bvwp-projekte.de/>

19.08.16 07:38:28 <https://duckduckgo.com/?q=b74-g10&atb=v19&ia=web>

19.08.16 07:38:31 <https://duckduckgo.com/?q=b74-g10+niedersachsen&atb=v19&ia=web>

19.08.16 07:38:54 https://r.duckduckgo.com//?kh=-1&uddg=http%3A%2F%2Fwww.bvwp-projekte.de%2Fstrasse%2Fdownload_plaene%2FNI%2FB74-G10-NI%2FLPL

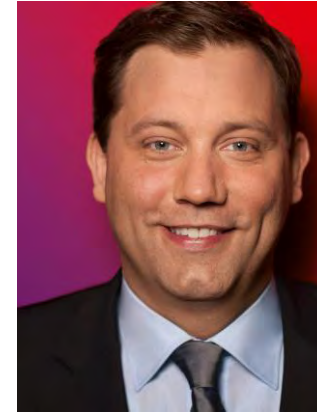
19.08.16 07:39:00 <http://bvwp-projekte.de/>



This is really bad to see something like this – especially if it is connected with my own name.

I don't know, why I was searching for "Tebonin" at that time.

More members of parliament and their employees





Employee of Helge Braun, CDU – Assistant Secretary of the German Chancellor

How we did it – the “hacking” part



Social engineering



Search for people, jobs, companies, and more...



Advanced



Home

Profile

My Network

Learning

Jobs

Interests

Business Services

Try Premium for free

Edit Background



Anna Rosenberg

Senior Consultant bei Meez Technology

Tel Aviv - Jaffa, Tel Aviv District, Israel | Marketing and Advertising

Current: Meez Technology

Education: Universität Hamburg

View profile as ▼

104
connections

[Update your public profile settings](#)

Contact Info

No, do not publish an update to my network about my profile changes.

No

Add a section to your profile – be discovered for your next career step.

technology meets creativity

DATA DRIVEN CONSULTING

LEARN MORE



WHAT WE DO

About

WHAT WE DO

PRODUCTS

technology meets creativity

PRODUCTS



[About](#)

[WHAT WE DO](#)
[PRODUCTS](#)

[CUSTOMIZED CAMPAIGNS](#)



CAREERS

About

WHAT WE DO
PRODUCTS

CAREER AT MEEZ

We are constantly looking for creative minds, analysts and dedicated doers for our team. We can

CONTACT US

If you are just as excited about the potential of technology and creativity, then we'd love to hear from you. We are a small, international team, based in beautiful and prospering Tel Aviv.

Mail: mail@meez.tech

What we have discovered



14 days (live) access

3 million (German) User Ids

Browsing data for one month

```
cat *.csv | grep "%40polizei.de"
```


< Zurück

Volkswagen Passat Variant ✕

von 500 € ✕

bis 17.500 € ✕

Erstzulassung von 2012 ✕

bis zu 70.000 km ✕

ab 103 kW ✕

Kraftstoff: Diesel ✕

Deutschland ✕

[Alle Filter entfernen](#)

PREIS

500 € ▾

17.500 € ▾

MwSt. ausweisbar

Erstzulassung

2012 ▾

bis ▾

Kilometerstand

von ▾

70.000 km ▾

Suche speichern

[Neue Angebote per E-Mail](#)



Volkswagen Passat Variant 2.0 TDI DSG Trendli... 

€ 14.777,- ** | 42.796 km | 09/2013 | 103 kW 140 PS

Gebraucht, Diesel

Sehr geehrte Damen und Herren, im Rahmen eines hier bearbeiteten **Ermittlungsverfahrens wegen Computerbetrug (Aktenzeichen)** benötige ich gem. § 113 TKG i.V.m. § 100j StPO eine Auskunft zu Bestandsdaten zu folgender **IP-Adresse: xxx.xxx.xxx.xxx** Zeitstempel: **xx.xx.2016, 10:05:31** MESZ

Die Daten werden für die Ermittlung des Täters benötigt. Bitte übersenden Sie Ihre Antwort per Email an die Adresse

Vorname.Nachname@polizei.Bundesland.de oder per Telefax.

Vorname Nachname

Kriminalhauptkommissar

Kriminalpolizeidirektion, **Ort**

CyberCrime

Telefonnummer

Ladies and Gentlemen, because of an **investigation concerning computer fraud** (file number), which I have dealt with here, § 113 TKG i.V.m. § 100j StPO I need information on following **IP address: xxx.xxx.xxx.xxx Time stamp: xx.xx.2016, 10:05:31 CEST**

The data is needed to identify the offender. Please send your answer by e-mail to the following address

Firstname.lastname@police.state.de or by fax.

first name

Last Name

Detective Chief Place of **county**

Cybercrime

phone number

Who did this



Browser Plugins

Neu registrieren oder anmelden | Andere Anwendungen mozilla

Add-ons

ERWEITERUNGEN THEMES SAMMLUNGEN MEHR...

Willkommen bei den Firefox-Add-ons. Wählen Sie aus Tausenden von Zusatzfunktionen und Stilen, um sich Firefox zu Eigen zu machen.

Neu & Jetzt

Holen Sie sich die neuesten Add-ons, die man jetzt haben muss.

- Instagram Video Do...
Download videos from Instagram (and as plus, easily save photos).
- I don't care about co...
Get rid of annoying cookie warnings from almost all 'infected' websites!
- Gmail™ Notifier +
Gmail™ Notifier Plus alerts you of new email messages. Users can also read or reply to email...

ERKUNDEN
Vorgestellt
Beliebtste
Am höchsten bewertet

Vorgestellte Erweiterungen

- Google search link fix
Datenschutz & Sicherheit (145)
- DownThemAll!
Download-Verwaltung (1.824)
- FindBar Tweak
- Download Statusbar

BELIEBTESTE

- Adblock Plus
18.474.289 Benutzer
- Video DownloadHelper
4.339.028 Benutzer
- Easy Screenshot
2.669.739 Benutzer

Chrome Web Store Anmelden

Im Chrome Web Store suchen

Vorgestellt

Best of 2016

Das Beste aus 2016
Vorgestellt
Erweiterungen des Jahres

SAMMLUNG ANZEIGEN

Erste Schritte

Mit diesen Erweiterungen können Sie Chrome optimal nutzen.

Alle anzeigen

- Office Online
★★★★ (904) KOSTENLOS
- Google Hangouts
★★★★ (29324) KOSTENLOS
- Chrome-Erweiterung für Google ...
★★★★ (1652) KOSTENLOS
- LastPass: Free Password Manager
★★★★ (19190) KOSTENLOS

https://chrome.google.com/webstore/category/collection/best-of-2016?utm_source=chrome-ntp-icon



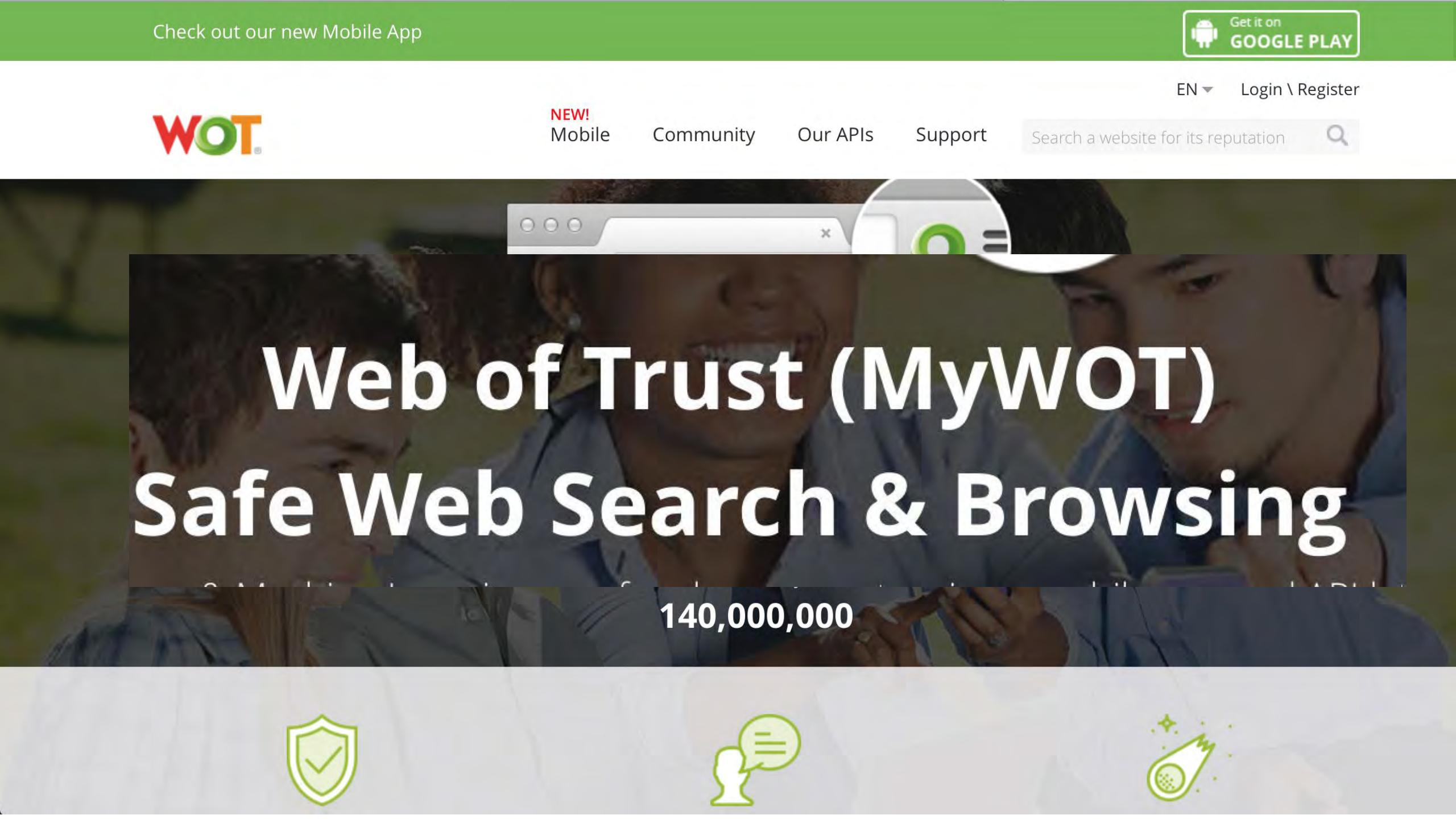
NEW!
Mobile

Community

Our APIs

Support

Search a website for its reputation



Web of Trust (MyWOT)

Safe Web Search & Browsing

140,000,000



[DATUM] 11:15:04 <http://what.kuketz.de/>

[...]

[DATUM] 15:49:27 <https://www.ebay-kleinanzeigen.de/p-anzeige-bearbeiten.html?adId=xxx>

[DATUM] 13:06:23 <http://what.kuketz.de/>

[...]

[DATUM] 11:22:18 <http://what.kuketz.de/>

[DATUM] 14:59:30 <http://blog.fefe.de/>

[...]

[DATUM] 14:59:36 <http://what.kuketz.de/>

[DATUM] 14:59:44

https://www.mywot.com/en/scorecard/what.kuketz.de?utm_source=addon&utm_content=rw-views

[...]

[DATUM] 13:48:24 <http://what.kuketz.de/>

[...]

test by Mike Kuketz / www.kuketz-blog.de



NEW!

Mobile

Community

Our APIs

Support

Search a website for its reputation



Privacy

Privacy policy

Privacy Policy

3. Automatically Collected Information – When you install and use the WOT app or WOT extension, we also automatically collect information from you related to your use of the WOT Services and your web browsing activity. This information may include, without limitation, the following:

- From WOT desktop/mobile extension and desktop app - Internet Protocol Address; device type; operating system and browser; web pages visited and time stamp of the visit; automatically-generated GUID and WOT user ID.

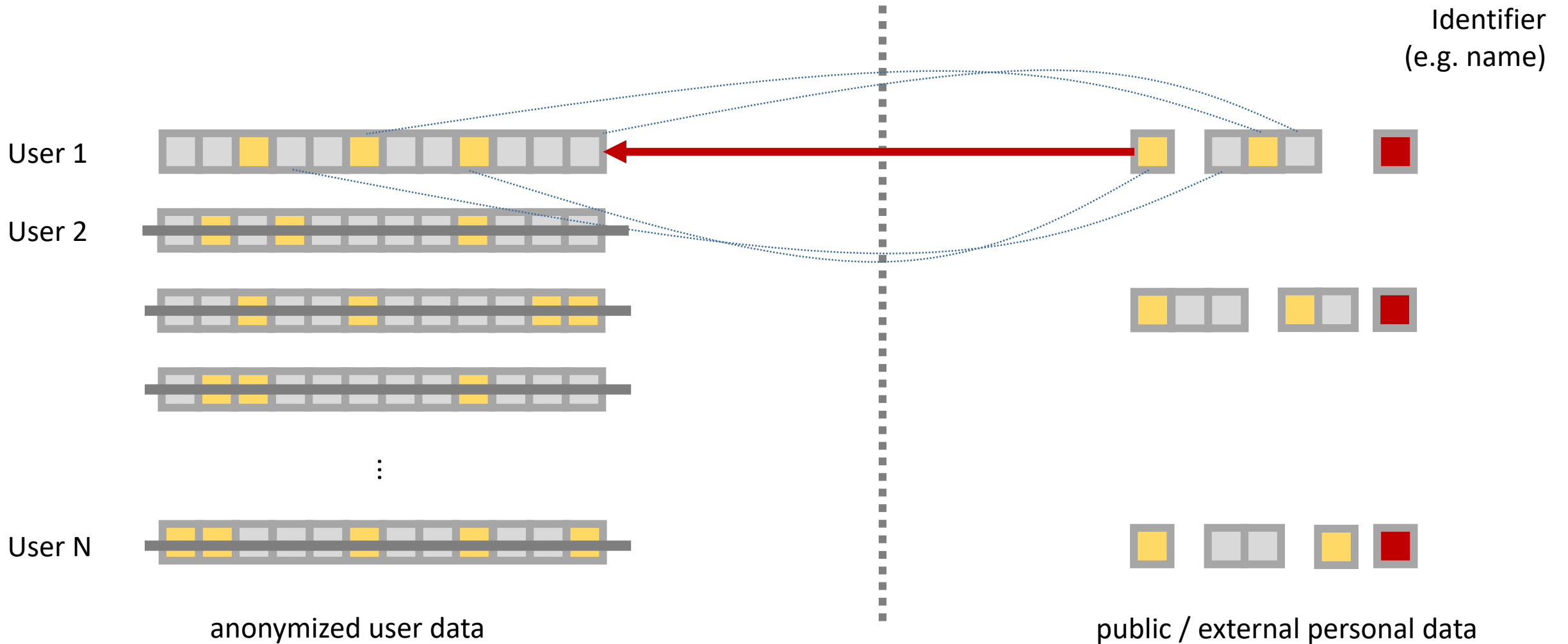
intended to make our users fully aware of the data we collect, the way it is stored, used, and shared, and our users' rights in relation to such data and practices. We greatly value our users and their privacy rights, and we encourage all of our users to read and become familiar with this

or de-identified information that is not collected together, or otherwise associated by us, with any Registration Information or other Personal Information, if any, that we hold about you. When we automatically collect information, we go to great lengths to make sure that Non-Personal Information remains anonymous, as further described below.

Information and Non-Personal Information in accordance with the terms of this Privacy Policy. **IF YOU DO NOT ACCEPT THE TERM OF THIS PRIVACY POLICY, PLEASE DO NOT ACCESS OR USE THE WOT PRODUCTS OR WOT SERVICES.**

- **Information We Collect**
- **How We Use the Data**
- **Sharing Collected Data with Third Parties**
- **Cookies**
- **Do Not Track Disclosure**
- **Your Responsibility for User-Generated Content**
- **Data Storage and Security**
- **Data Retention**

How does deanonymization work?



"Instant" deanonymization via unique URL

The image shows two browser screenshots side-by-side, illustrating how a unique URL can deanonymize a user's profile across different platforms.

Top Screenshot (Twitter Analytics):
The browser address bar shows the URL: `https://analytics.twitter.com/user/japh44/home`. The page header includes the Twitter logo, "Analytics", and navigation links: "Startseite", "Tweets", "Zielgruppen", "Ereignisse", and "Mehr". The user's name "Andreas Dewes" and a dropdown arrow are visible in the top right. The main content area features a profile picture of a man with glasses, the text "Mein Account", and the name "Andreas Dewes @japh44". A "2D" badge is present in the top right corner of the profile banner.

Bottom Screenshot (XING AG [DE]):
The browser address bar shows the URL: `https://www.xing.com/profile/Andreas_Dewes?sc_o=mx_b_p`. The XING logo is in the top left. A search bar contains the text "Suchen Sie nach Jobs, Kontakten, Events ...". To the right of the search bar are links for "Erweiterte Suche", "Neue Kontakte finden", and "Hilfe". The profile header shows "Dr. Andreas Dewes" and a notification: "Jetzt Premium-Mitglied werden! Mehr Infos". Below the header, there is a "PREMIUM Profilbesucher" section with a row of profile pictures. On the right, there is a "Profilansicht" section with an "Editiermodus" dropdown menu.

Combinatorial deanonymization

https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

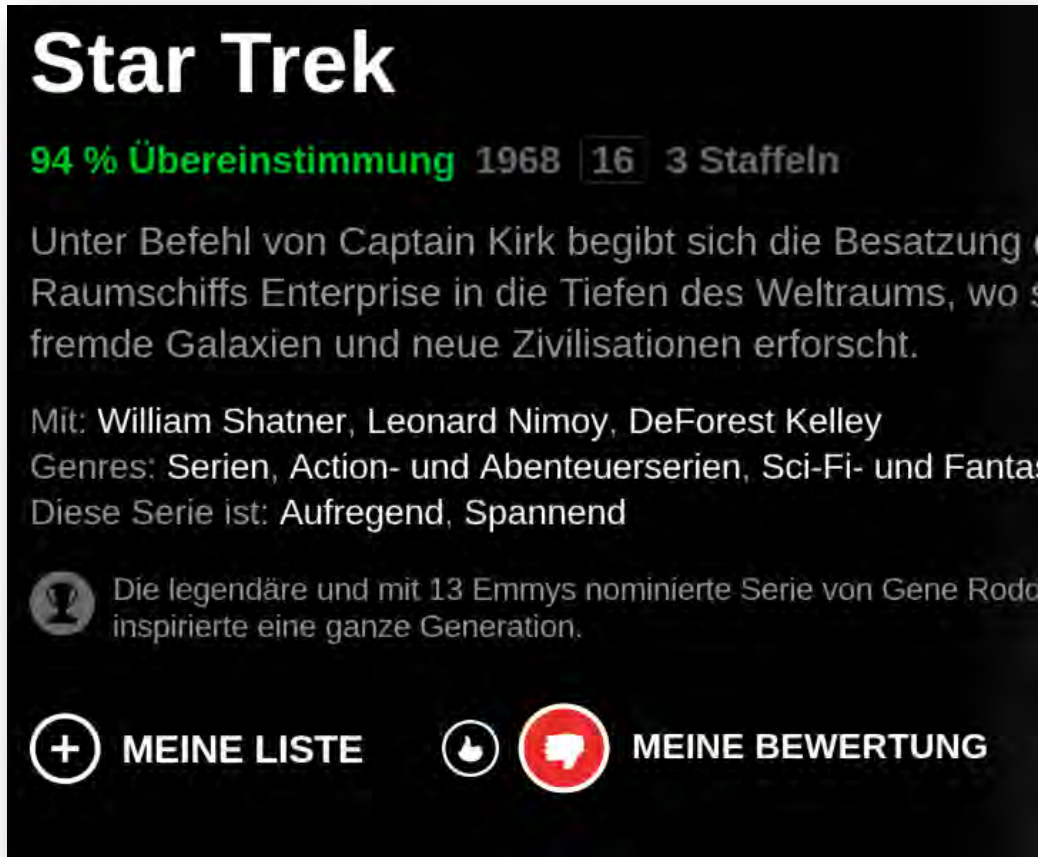
We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

[cs.CR] 22 Nov 2007

Netflix Data vs. IMDB Data

Provided anonymized



Star Trek
94 % Übereinstimmung 1968 16 3 Staffeln
Unter Befehl von Captain Kirk begibt sich die Besatzung des Raumschiffs Enterprise in die Tiefen des Weltraums, wo sie fremde Galaxien und neue Zivilisationen erforscht.
Mit: William Shatner, Leonard Nimoy, DeForest Kelley
Genres: Serien, Action- und Abenteuerserien, Sci-Fi- und Fantasy
Diese Serie ist: Aufregend, Spannend
Die legendäre und mit 13 Emmys nominierte Serie von Gene Roddenberry inspirierte eine ganze Generation.
MEINE LISTE MEINE BEWERTUNG

ratings associated with user name / real name



1 out of 16 people found the following review useful:
This show is pathetic!
★★★★★
Author: [redacted]
20 October 2013
*** This review may contain spoilers ***
Lets just sum it up. One character has basically one line for every stupid episode "dead Jim!".
Virtually every single episode they are on different planet. So that means the writers of this show that meant to recycle Nazi's, Roman's, Greek's, God and wants to scream ENOUGH!

Our data set



URLs
(insufficiently anonymized)



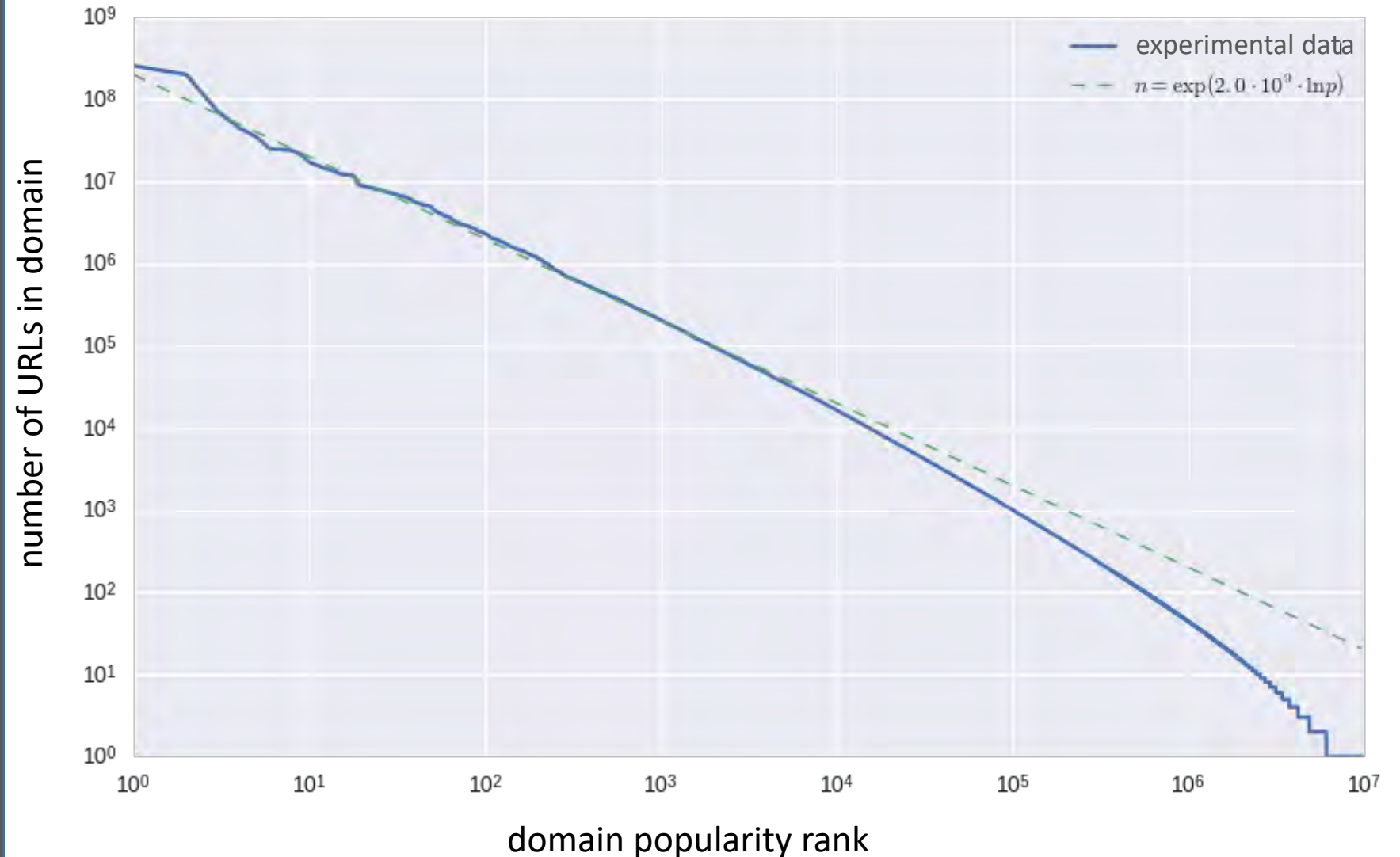
Domains



Users

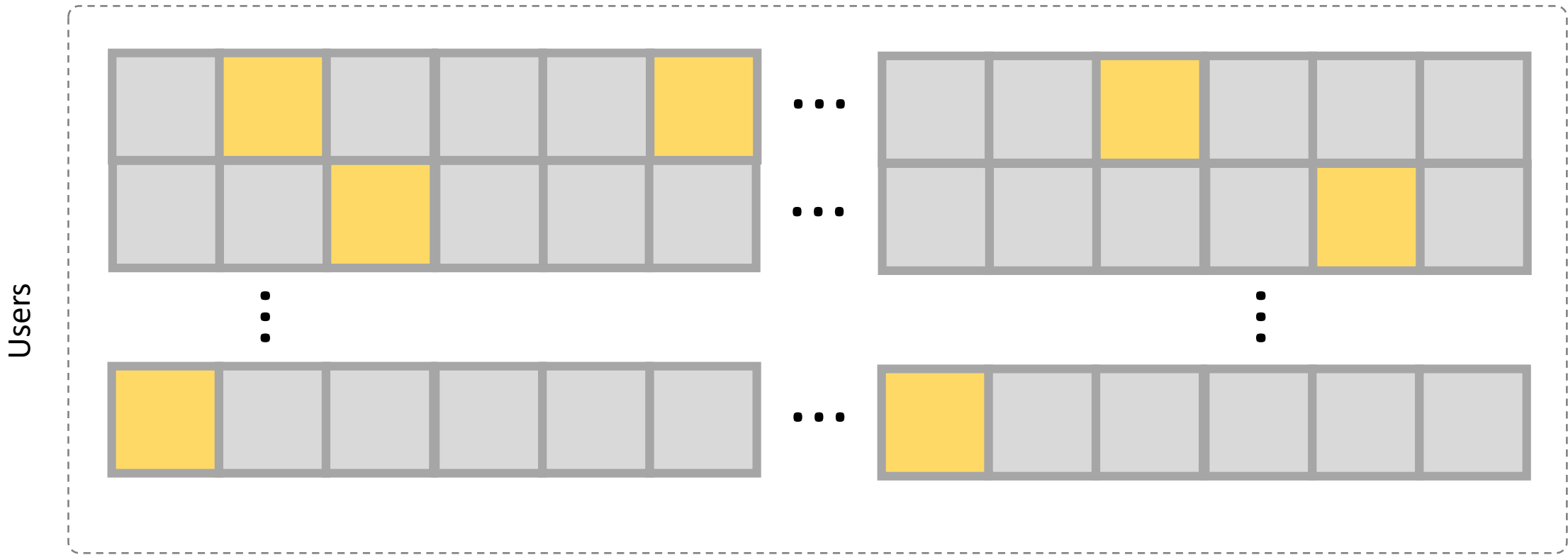
- We remove everything but the **domain** and user Id
- „Did this user visit this domain?“ (yes / no)
- We investigate how easy it is to reidentify a user given his/her domain data
- We only look at users that have visited at least ten domains

Frequency analysis of domains



Let's categorize our users

Domains



=> sparsely populated matrix with **9.000.000** x **1.000.000** entries

Algorithm

- Generate user/domain matrix M
- Generate vector v with information about visited domains
- Multiply $M \cdot v$
- Look for best match

$$M = (...)$$

$$w = M \cdot v$$

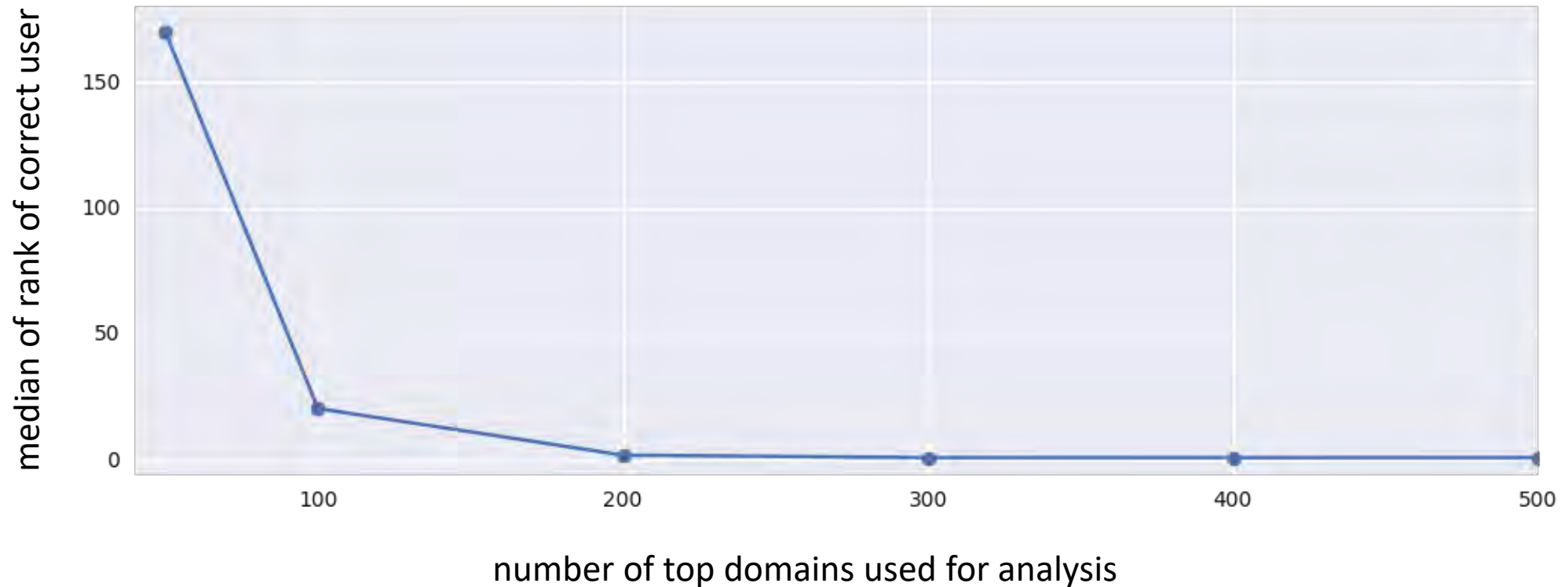
$$i = \operatorname{argmax}(w)$$

How unique am I?



How well does this work?

Top-200 domains are already sufficient to identify a large fraction of our users

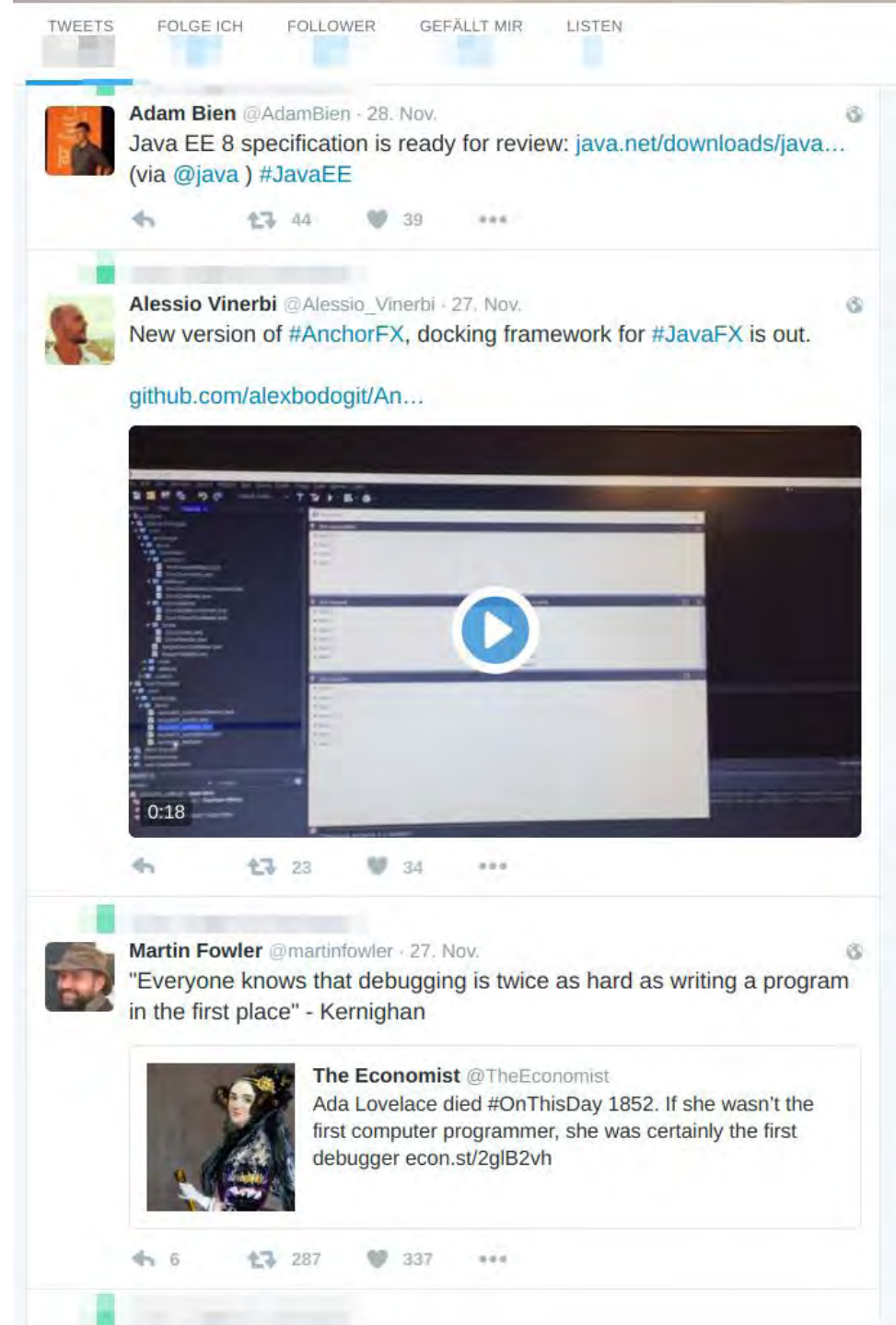


But how can public information be extracted?

Three examples

Twitter

- We use the Twitter API to download tweets from the relevant time period (one month)
- We extract URLs from the tweets and generate the associated domain by following the links
- We feed the domain information into our algorithm



The screenshot shows a Twitter interface with three tweets. At the top, navigation options are visible: TWEETS, FOLGE ICH, FOLLOWER, GEFÄLLT MIR, and LISTEN. The first tweet is from Adam Bien (@AdamBien) dated 28. Nov., mentioning the Java EE 8 specification. The second tweet is from Alessio Vinerbi (@Alessio_Vinerbi) dated 27. Nov., announcing a new version of AnchorFX and including a video player showing a code editor. The third tweet is from Martin Fowler (@martinfowler) dated 27. Nov., quoting Kernighan and featuring a tweet retweet from The Economist about Ada Lovelace.

TWEETS FOLGE ICH FOLLOWER GEFÄLLT MIR LISTEN

Adam Bien @AdamBien · 28. Nov.
Java EE 8 specification is ready for review: java.net/downloads/java...
(via @java) #JavaEE

44 39

Alessio Vinerbi @Alessio_Vinerbi · 27. Nov.
New version of #AnchorFX, docking framework for #JavaFX is out.
github.com/alexbodogit/An...

0:18

23 34

Martin Fowler @martinfowler · 27. Nov.
"Everyone knows that debugging is twice as hard as writing a program in the first place" - Kernighan

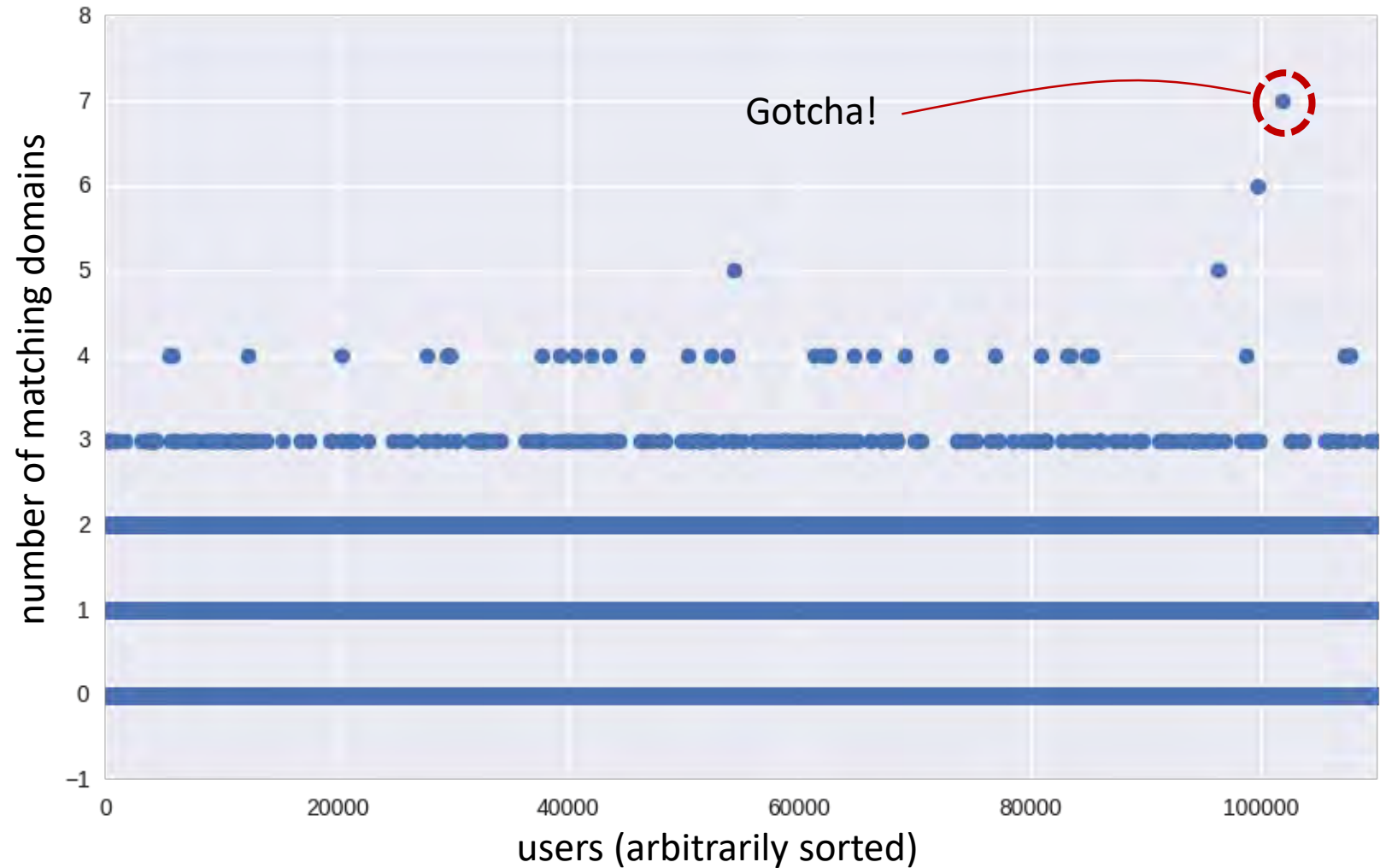
The Economist @TheEconomist
Ada Lovelace died #OnThisDay 1852. If she wasn't the first computer programmer, she was certainly the first debugger econ.st/2glB2vh

6 287 337

Visited Websites

github.com (2.584.681)
www.change.org (124.152)
fxexperience.com (394)
community.oracle.com (5161)
paper.li (2689)
javarevisited.blogspot.de (525)
www.adam-bien.com (365)
rterp.wordpress.com (129)

Examples

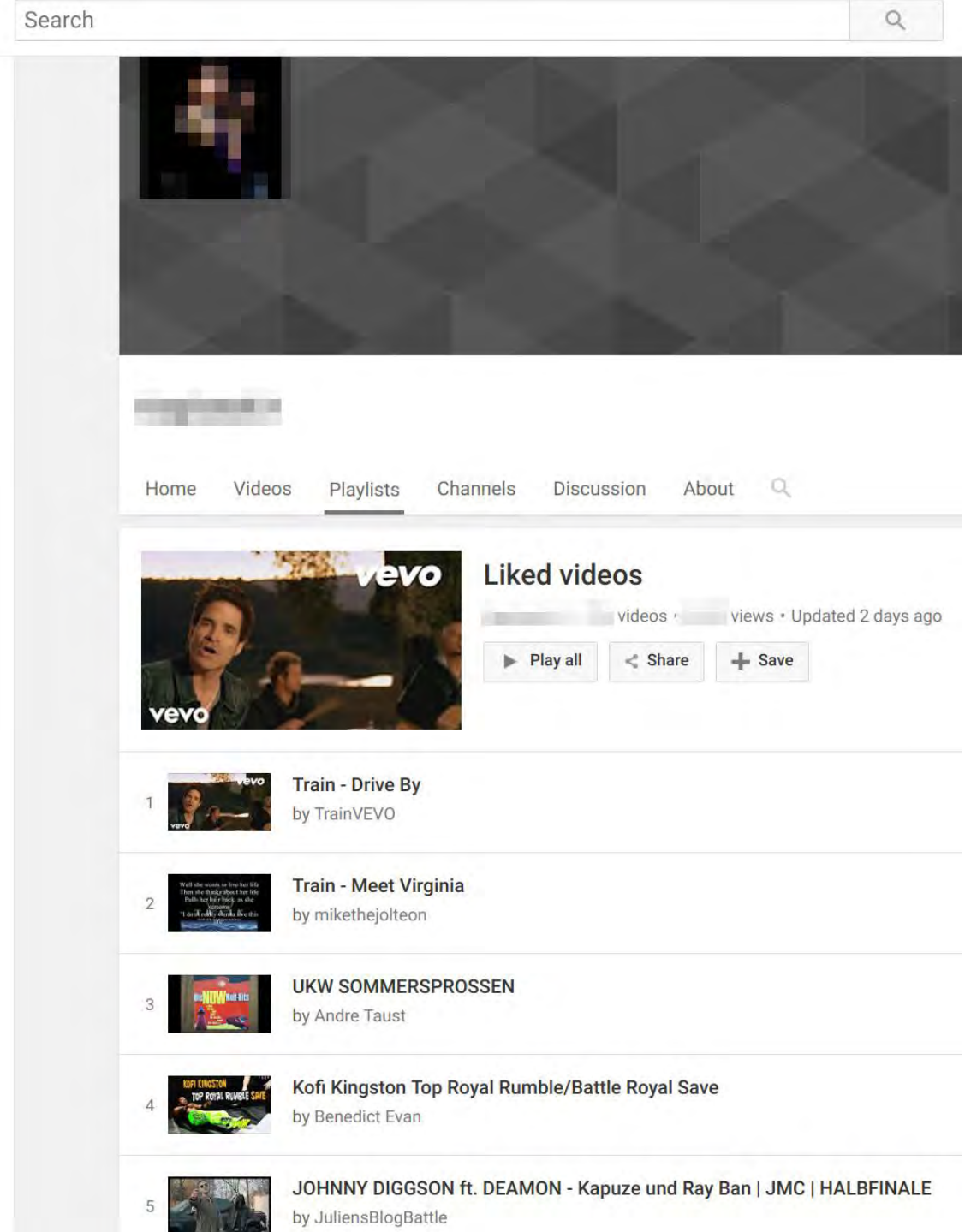


Seemingly harmless identifiers can betray you

[https://www.youtube.com/watch?v=DLzxrzFCy0s](https://www.youtube.com/watch?v=<u>DLzxrzFCy0s</u>)

Youtube

- We download public playlists from users (often linked via Google+)
- We extract the video IDs using the Youtube API
- We feed the resulting (full) URLs into our algorithm (this time with full URL info)

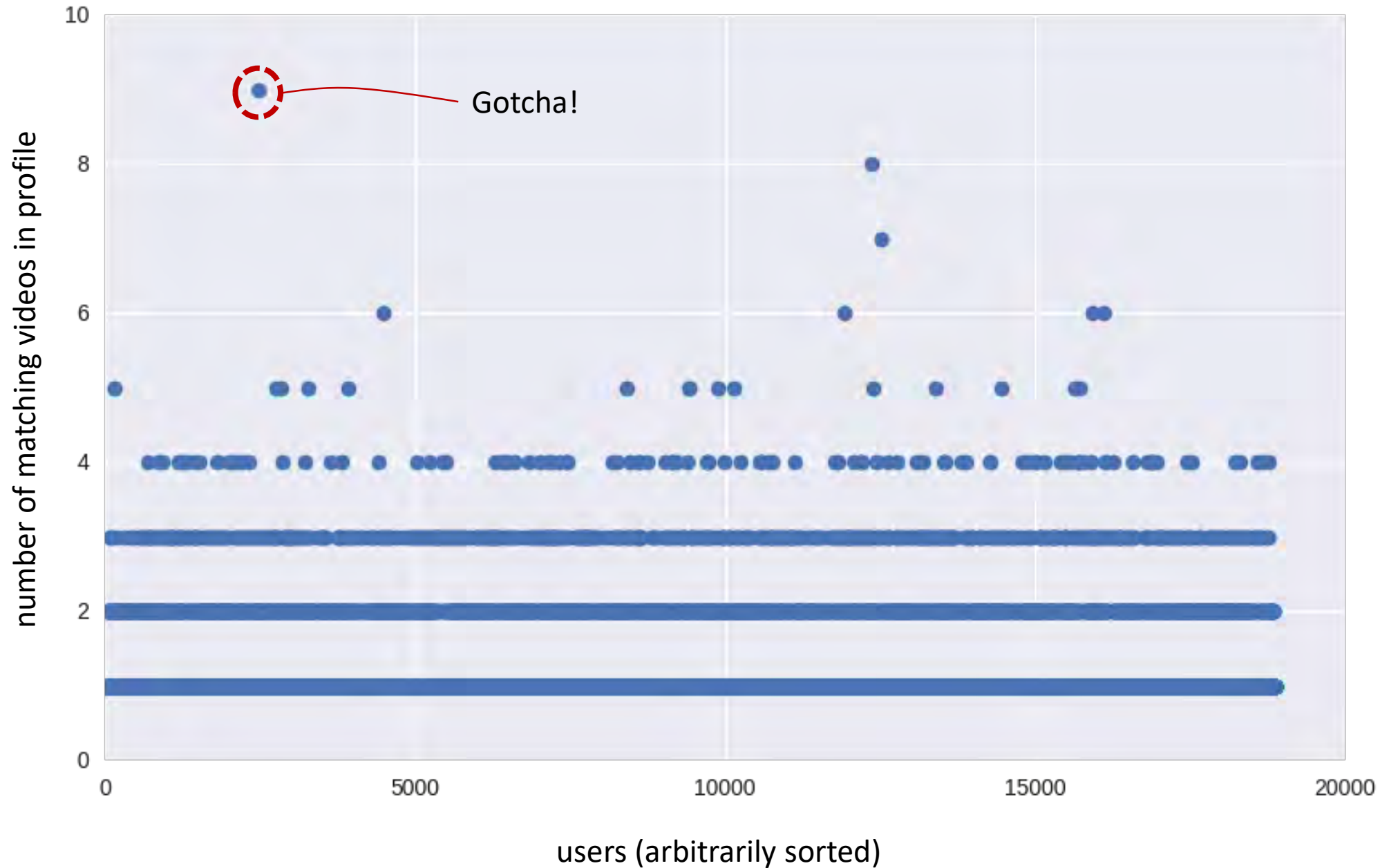


The image shows a screenshot of a YouTube channel's 'Liked videos' page. At the top, there is a search bar with the text 'Search' and a magnifying glass icon. Below the search bar is a large, dark, blurred video player area. Underneath the video player is a navigation bar with tabs for 'Home', 'Videos', 'Playlists', 'Channels', 'Discussion', and 'About', with a search icon on the right. The 'Playlists' tab is currently selected. Below the navigation bar, the 'Liked videos' section is displayed. It features a header with the text 'Liked videos' and a sub-header showing the number of videos and views, along with the text 'Updated 2 days ago'. Below the header are three buttons: 'Play all', 'Share', and 'Save'. The main content area shows a list of five liked videos, each with a thumbnail, a title, and the uploader's name. The videos are: 1. 'Train - Drive By' by TrainVEVO; 2. 'Train - Meet Virginia' by mikethejoltion; 3. 'UKW SOMMERSPROSSEN' by Andre Taust; 4. 'Kofi Kingston Top Royal Rumble/Battle Royal Save' by Benedict Evan; 5. 'JOHNNY DIGGSON ft. DEAMON - Kapuze und Ray Ban | JMC | HALBFINALE' by JuliensBlogBattle.

Example

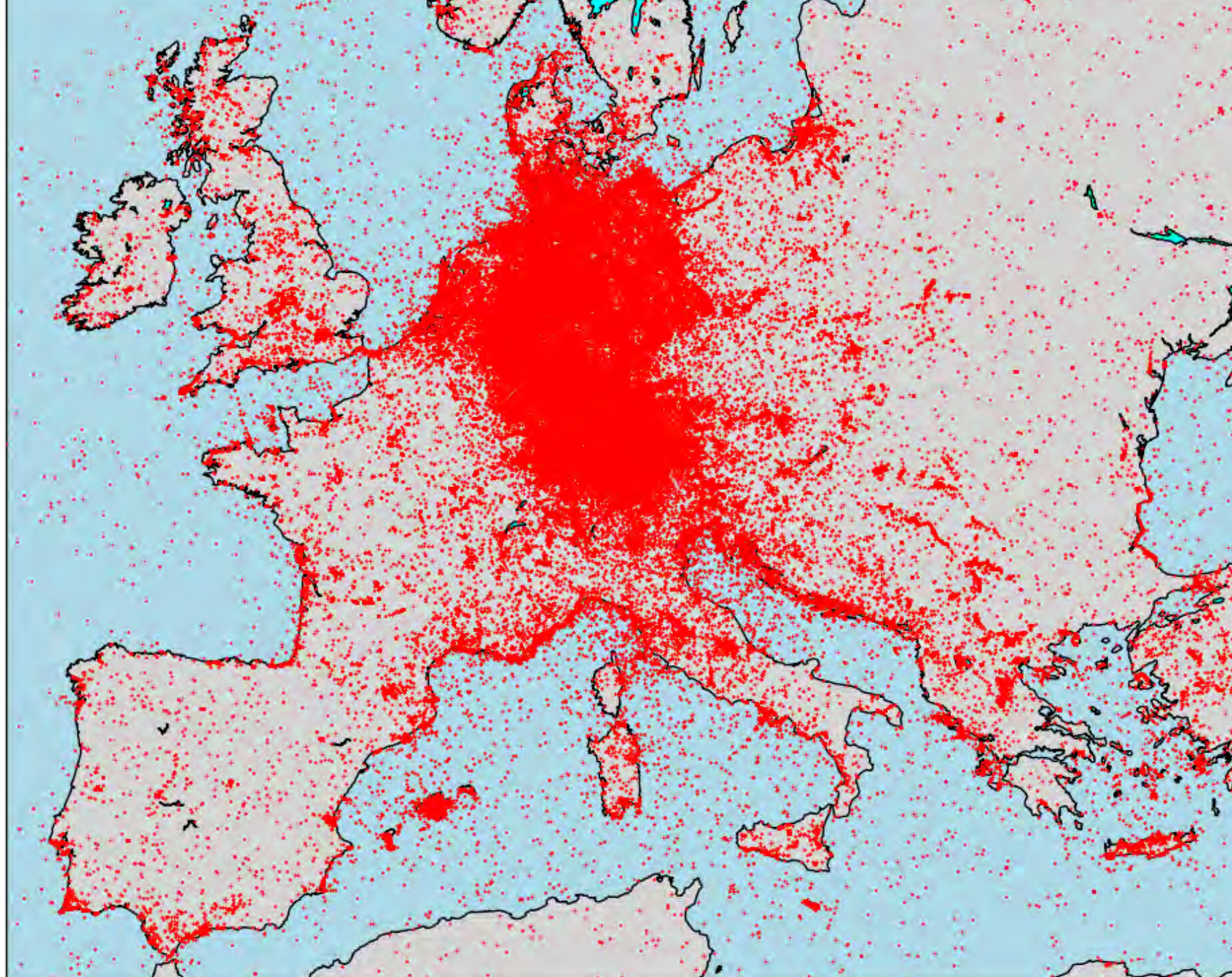
Video-IDs:

02Zm-Ayv-PA
18rBn4heThI
2ips2mM7Zqw
2wUv1TUi8kQ
34Na4j8AVgA
3VVuMIB2hC0
4fXvJHrbUTA
4u1aGjwiIbo
5BzkbSq7pww
5RDSkR8_AQ0
680R1Gq2YYU
6IHq9yv_qis
8d5QEwdHchk
...



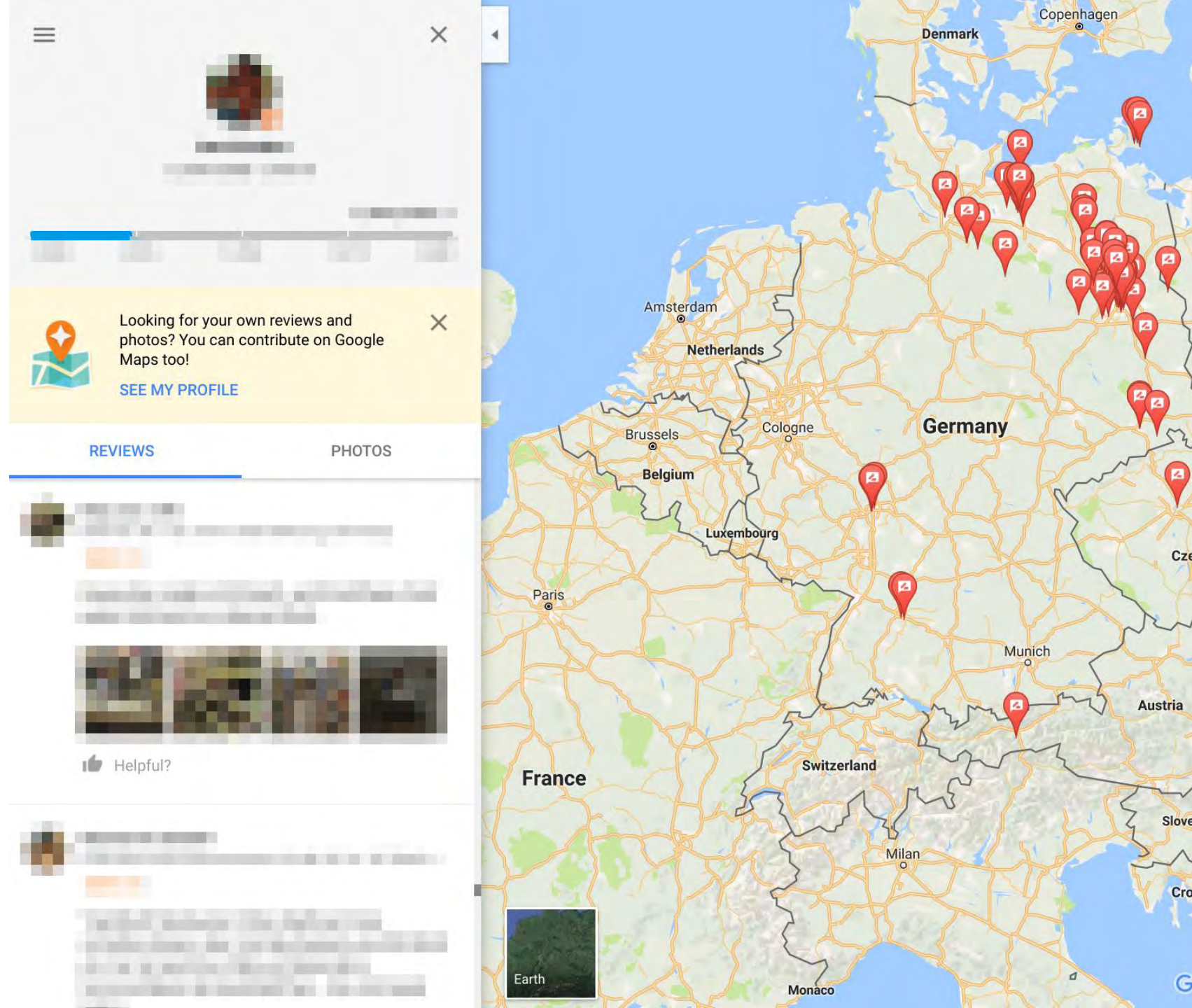
Geo-based identification

- We extract geo-data from Google Maps URLs (i.e. what coordinate was the user looking at)



Google Maps

- Ratings and photos are often publicly available (thanks again, Google+)
- Locations of interest could also be extracted from social media accounts
- A few data points are already enough to identify you



Can I hide in my data by generating noise?

(e.g. via random page visits)

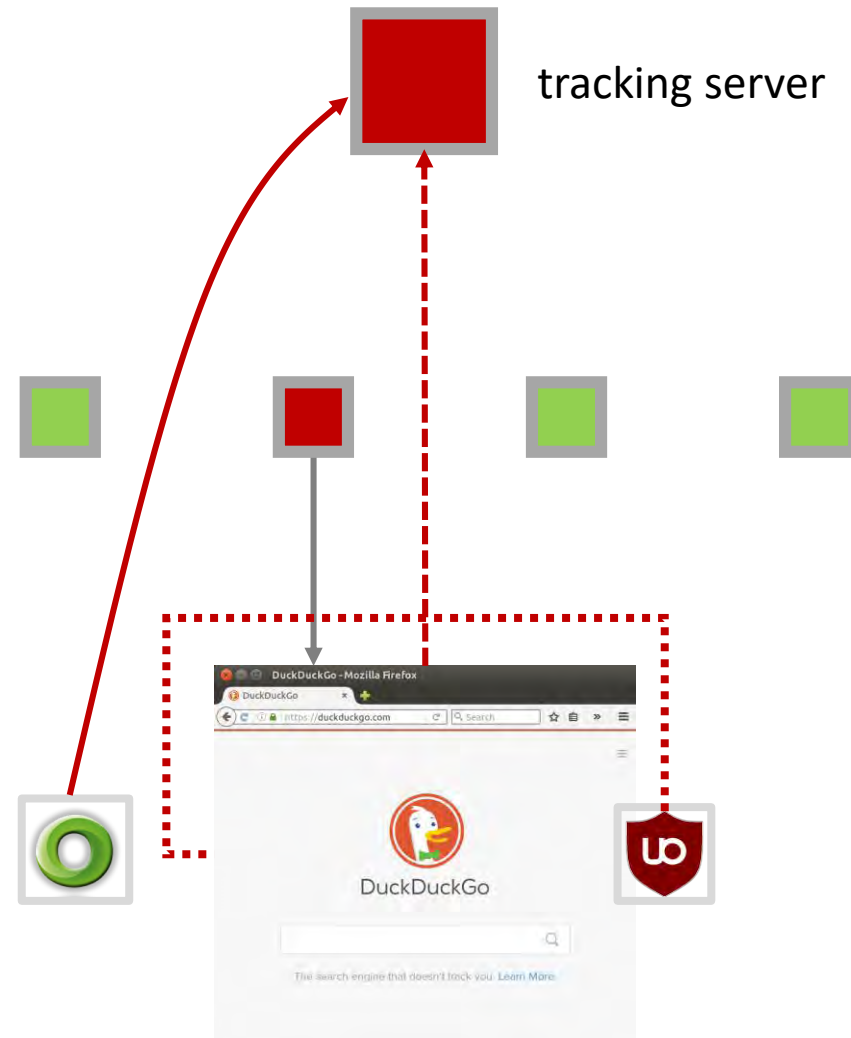
$\sqrt{(v)}$

Usually not

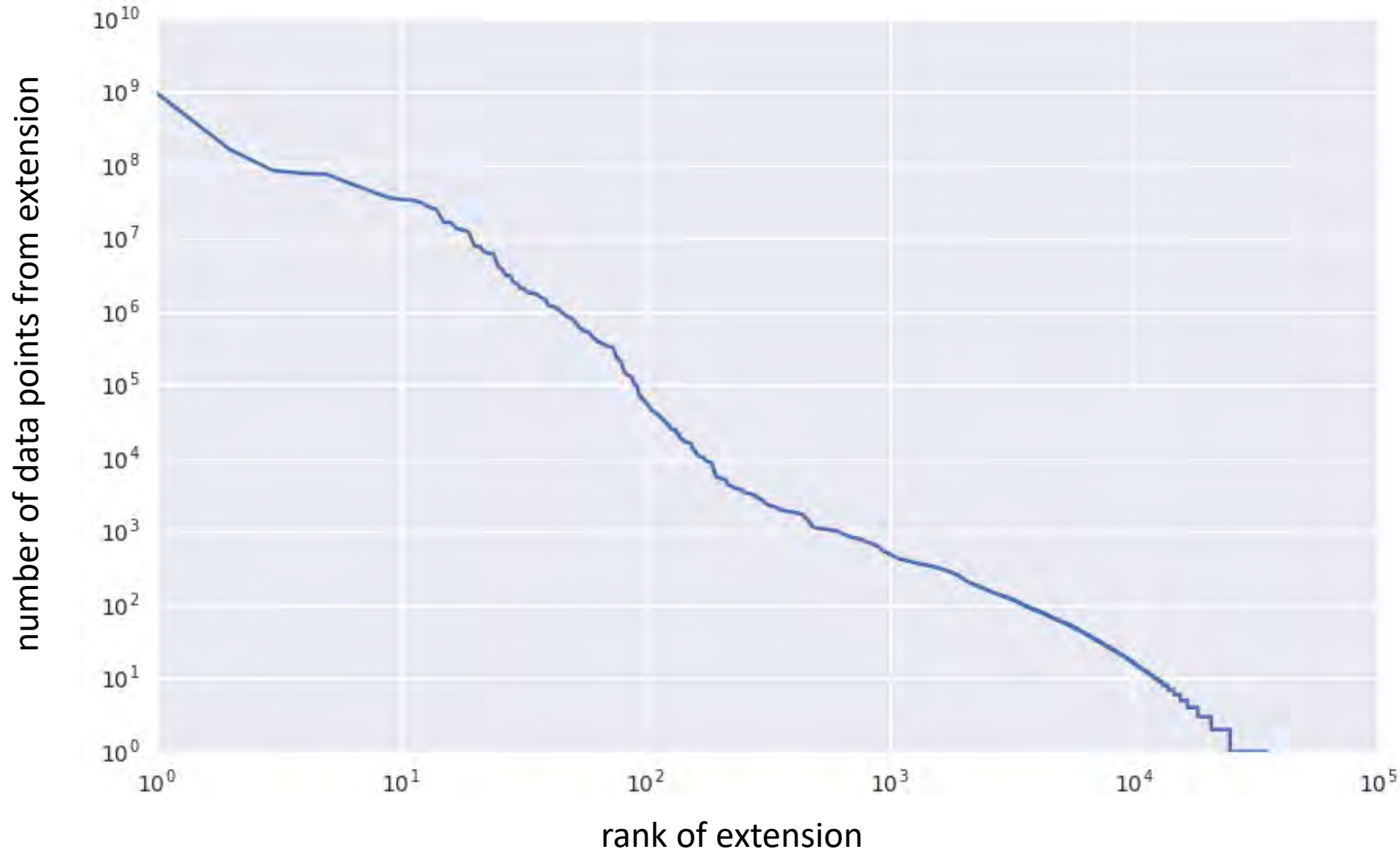


$\mathit{argmax} \|M \cdot v\|$ is robust against isolated (additive) perturbation

Why use extensions for tracking?



Analysis of data points per extension



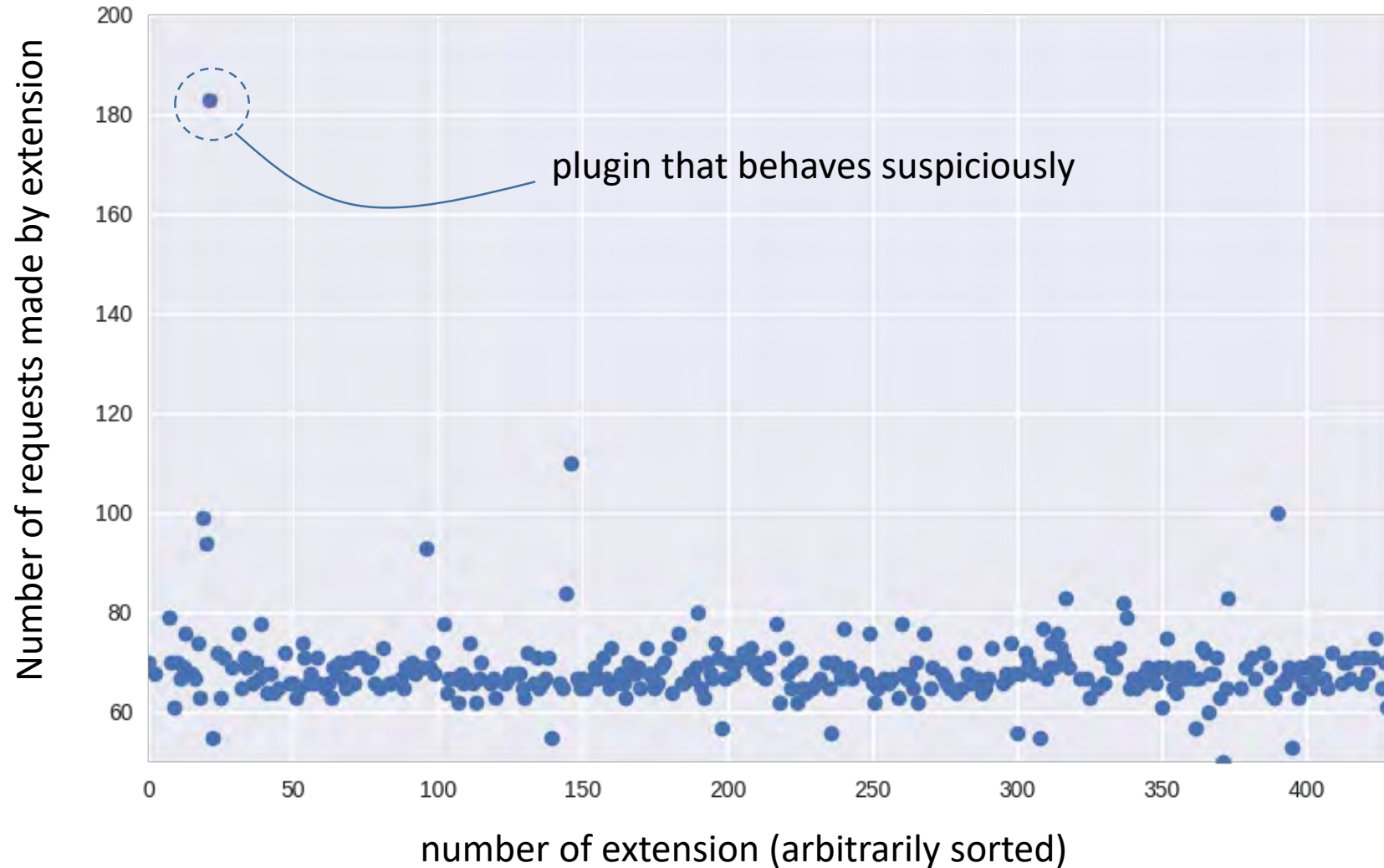
95 % of the data comes from only **10 extensions**.

Many more are spying on their users, but have a small installation base.

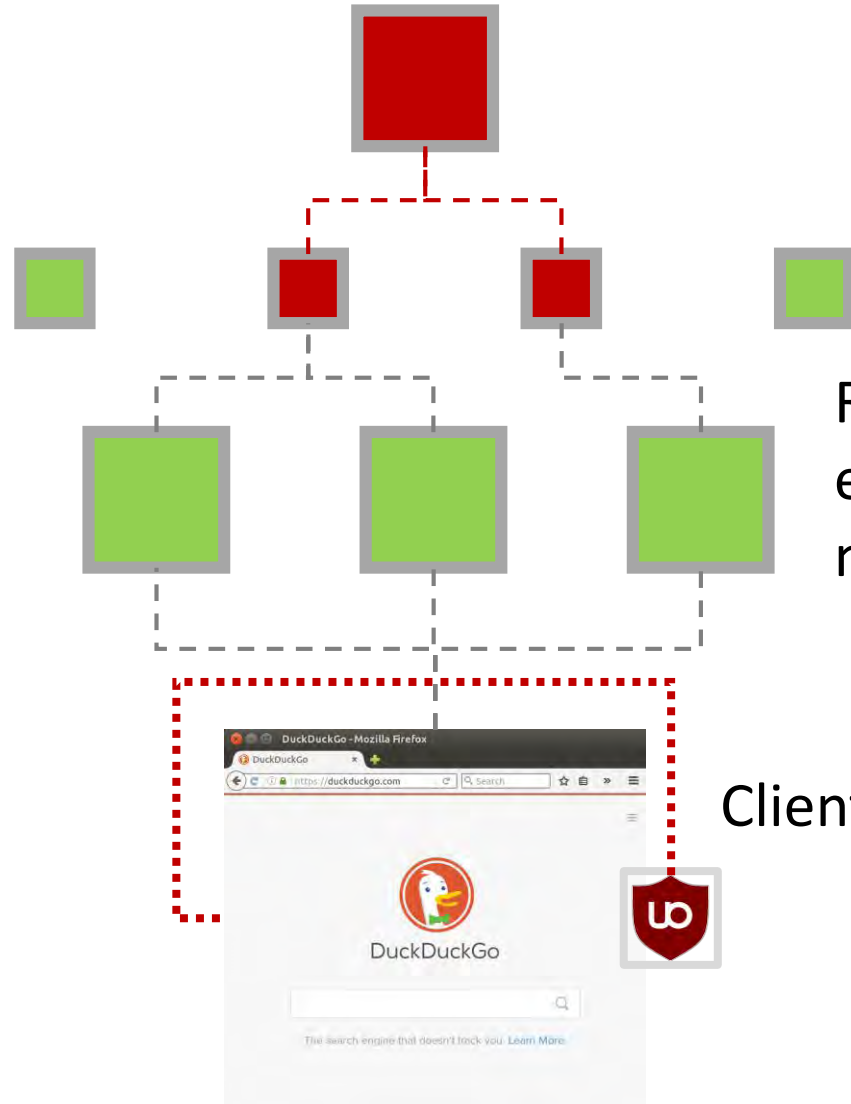
Up to **10.000 extension versions** affected (upper bound analysis via extension ID)

Behavior analysis of chrome extensions

(via Selenium Webdriver + Docker)



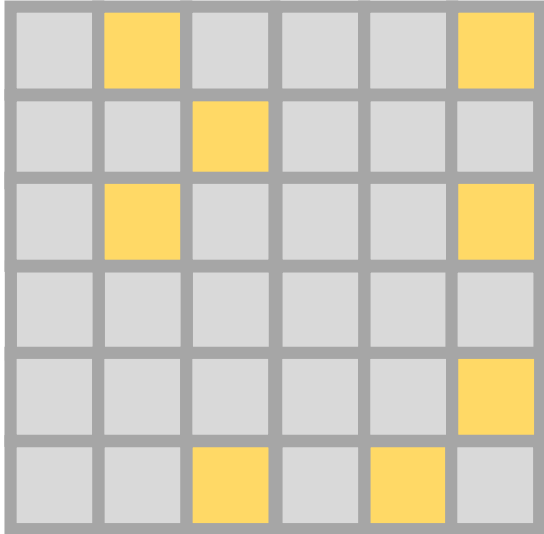
(How) can I protect myself?



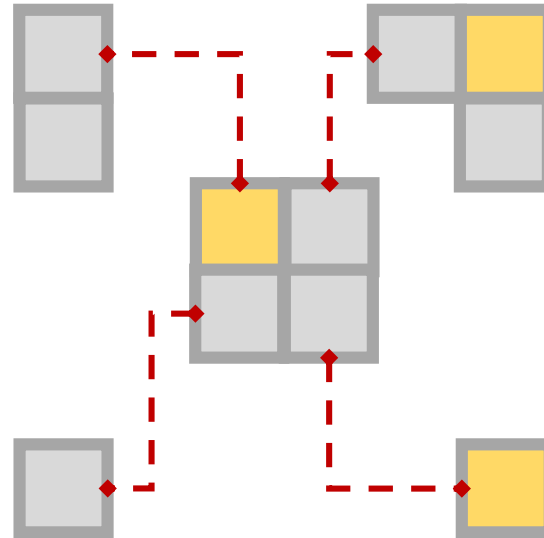
Rotating proxy servers ($n \gg 1$)
e.g. TOR or a VPN with rotating exit nodes

Client-side blocking of trackers

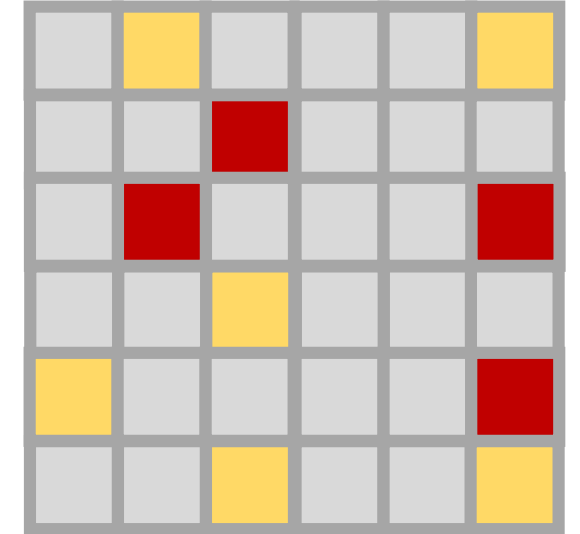
Takeaways



High-dimensional, user-related data is really hard to robustly anonymize (even if you really try to do so).



The increase in publicly available information on many people makes de-anonymization via linkage attacks easier than ever before.



Often, only a few external data points (<10) are sufficient to uniquely identify a person.

Special thanks to

Kian Badrnejad, NDR

Jasmin Klofta, NDR

Jan Lukas Strozyk, NDR

Martin Fuchs @wahlbeobachter

Stefanie Helbig

Mike Kuketz, kuketz-blog.de

Many anonymous sources and contributors

TV shows ARD Panorama, Panorama3 und ZAPP

<http://daserste.ndr.de/panorama/archiv/2016/Nackt-im-Netz-Intime-Details-von-Politikern-im-Handel,nacktimnetz110.html>