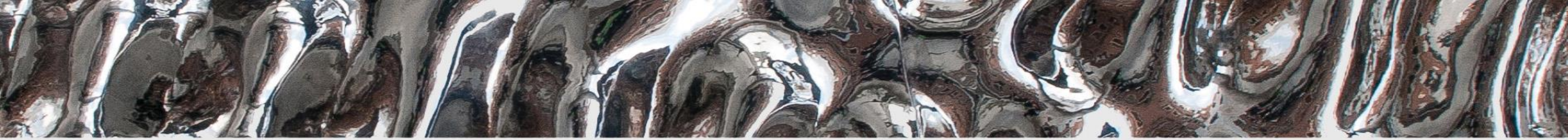


UiO : Department of Informatics
University of Oslo

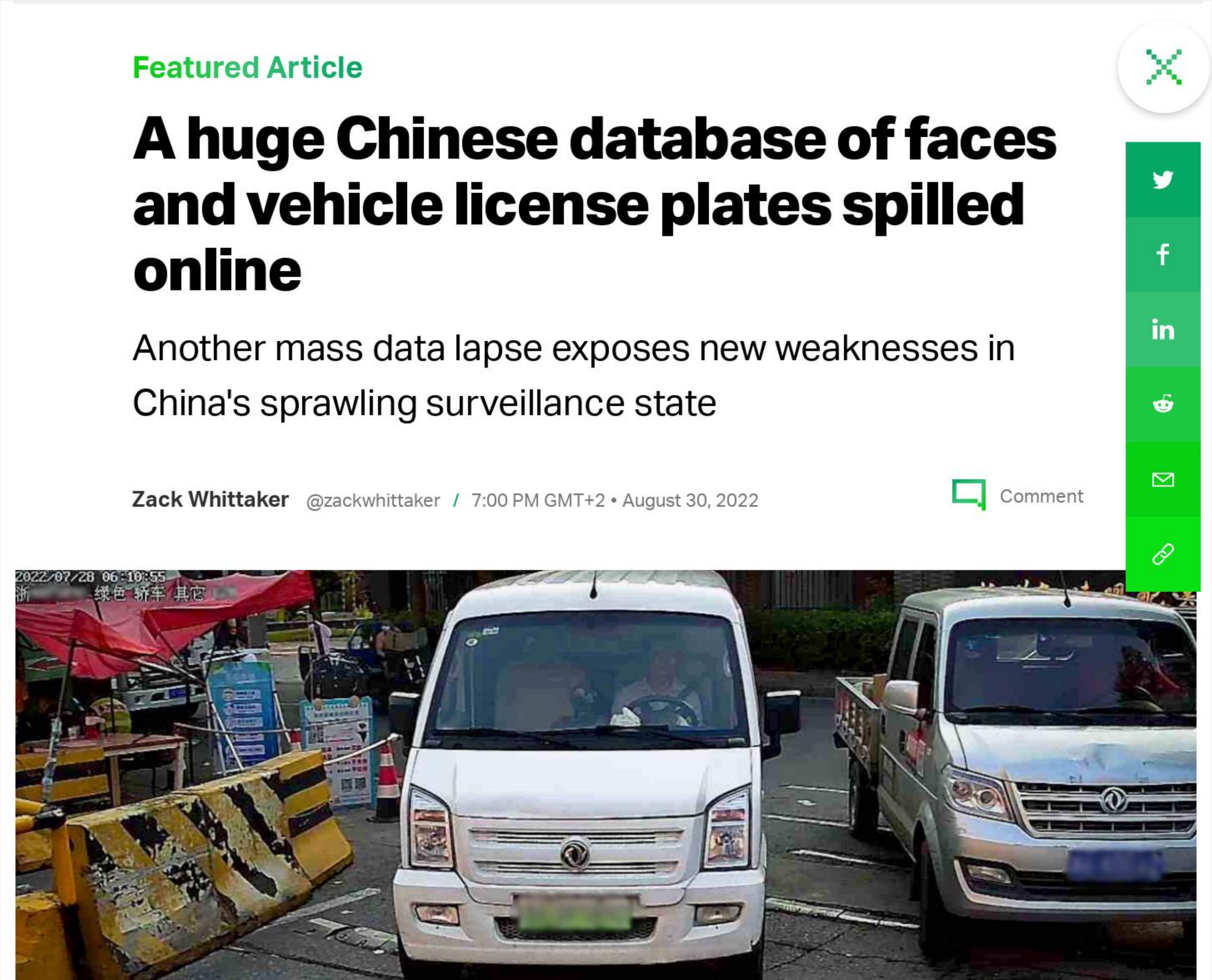
Privacy & Sikt

IN5000

14 March 2023

Trenton Schulz

Why do people doing research in informatics need to know things about privacy?



Featured Article

A huge Chinese database of faces and vehicle license plates spilled online

Another mass data lapse exposes new weaknesses in China's sprawling surveillance state

Zack Whittaker @zackwhittaker / 7:00 PM GMT+2 • August 30, 2022

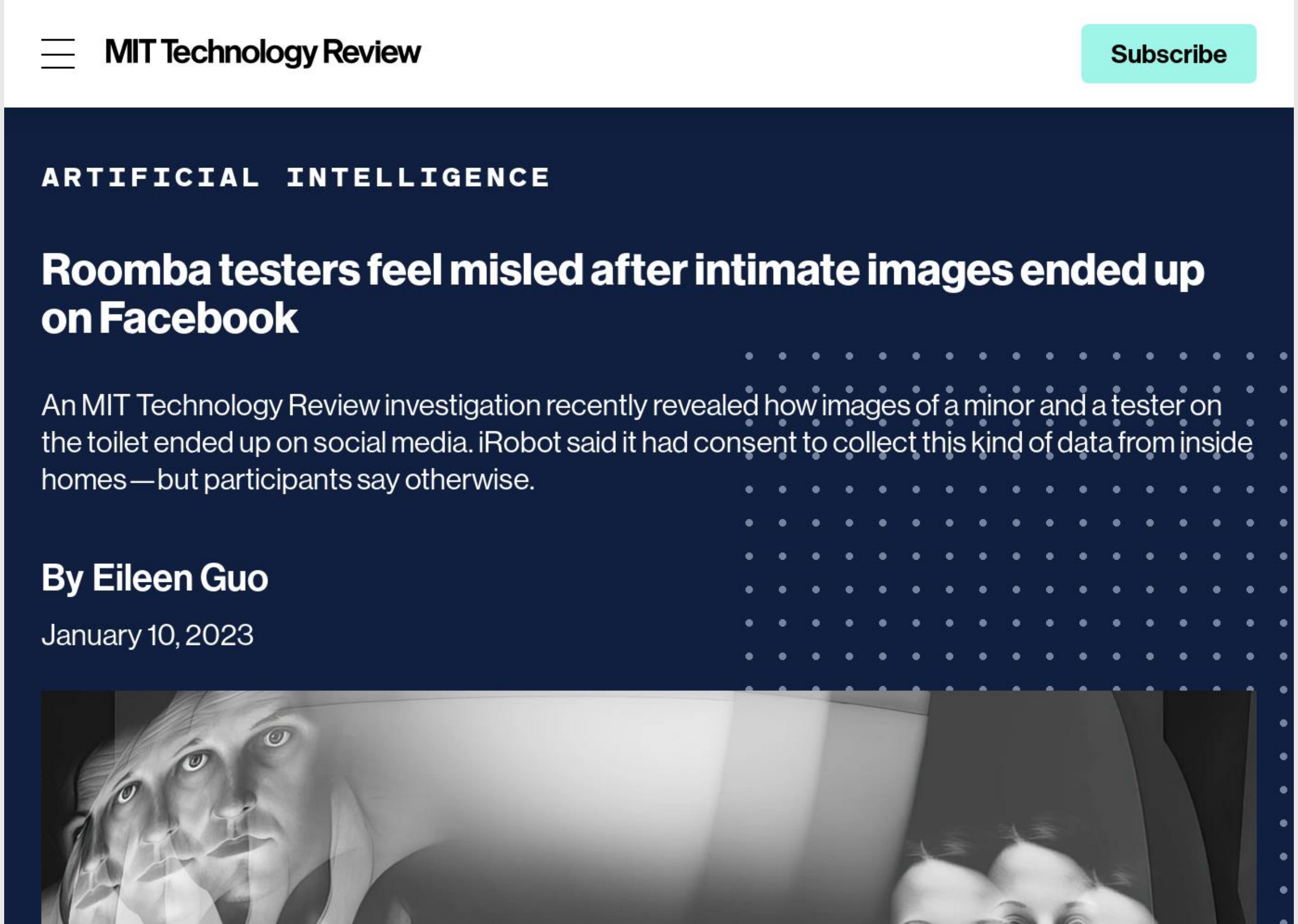
Source:
Techcrunch
30 August 2022
https://techcrunch.com/2022/08/30/china-database-face-recognition/

Data leaks happen, the most interesting thing in this case was the amount of data that was leaked.

> “At its peak the database held over 800 million records, representing one of the biggest known data security lapses of the year by scale, second to a massive data leak of 1 billion records from a Shanghai police database in June. In both cases, the data was likely exposed inadvertently and as a result of human error.”

Source:
Techcrunch
30 August 2022
https://techcrunch.com/2022/08/30/china-database-face-recognition/

Collecting personal data requires that people are informed, but making this clear and understandable is not always easy



ARTIFICIAL INTELLIGENCE

Roomba testers feel misled after intimate images ended up on Facebook

An MIT Technology Review investigation recently revealed how images of a minor and a tester on the toilet ended up on social media. iRobot said it had consent to collect this kind of data from inside homes—but participants say otherwise.

By Eileen Guo

January 10, 2023

Source:
MIT Technology Review
10 January 2023
https://www.technologyreview.com/2023/01/10/1066500/roomba-irobot-robot-vacuum-beta-product-testers-consent-agreement-misled/

Collecting personal data requires that people are informed, but making this clear and understandable is not always easy

"When MIT Technology Review reached out to iRobot for comment on the set of 15 images last fall, the company emphasized that each image had a corresponding consent agreement. **It would not, however, share the agreements with us**, citing 'legal reasons.' Instead, the company said the agreement required an '**acknowledgment that video and images are being captured during cleaning jobs**' and that '**the agreement encourages paid data collectors to remove anything they deem sensitive from any space the robot operates in, including children**.'"

Source:
MIT Technology Review
10 January 2023
https://www.technologyreview.com/2023/01/10/1066500/roomba-irobot-robot-vacuum-beta-product-testers-consent-agreement-misled/

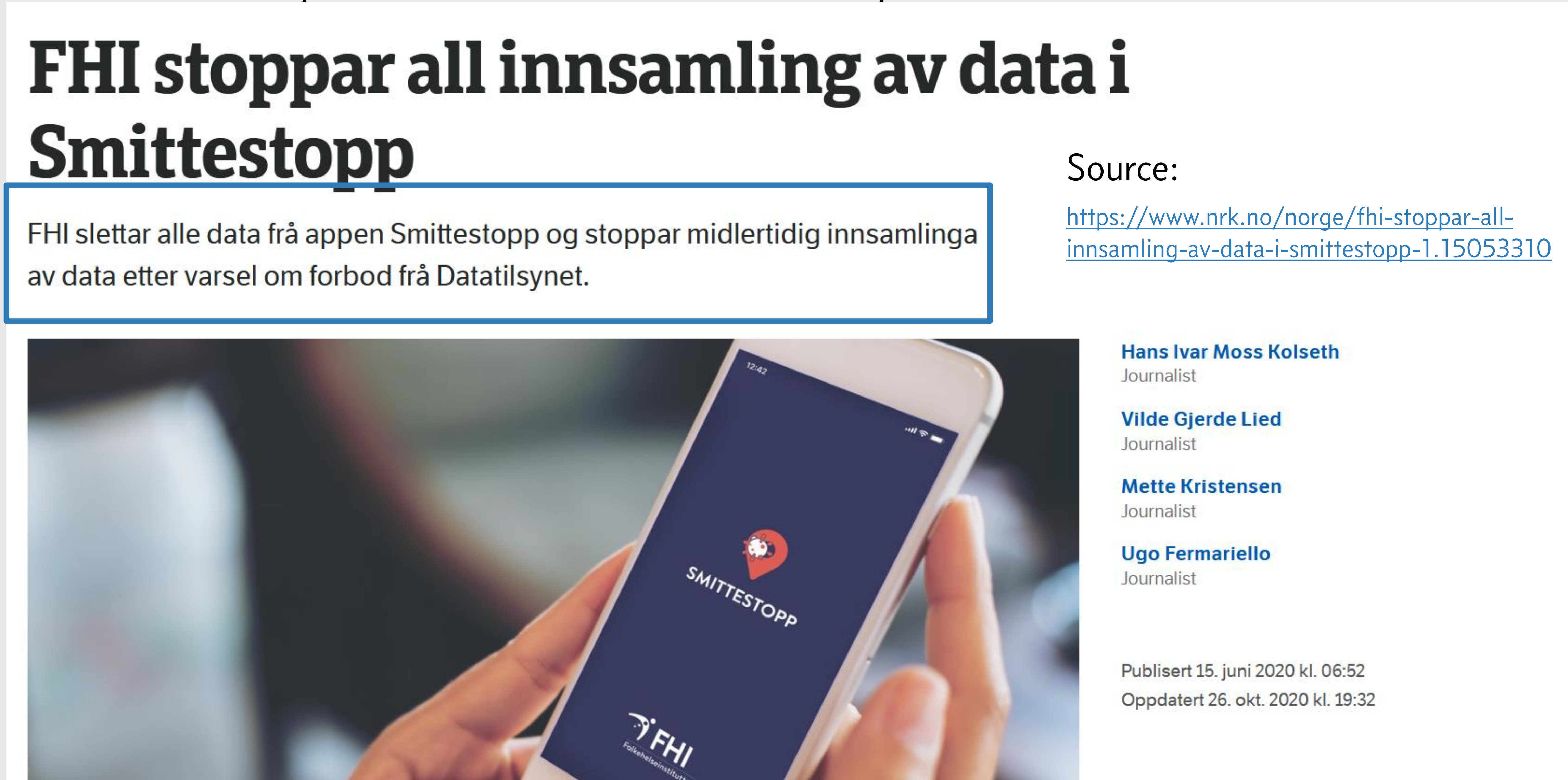
But upon MIT Technology Review reading the agreement

"The forms do contain the language iRobot previously laid out, while also spelling out the company's own commitments on data protection for test users. **But they provide little clarity on what exactly that means, especially how the company will handle user data after it's collected and whom the data will be shared with**."

Source:
MIT Technology Review
10 January 2023
https://www.technologyreview.com/2023/01/10/1066500/roomba-irobot-robot-vacuum-beta-product-testers-consent-agreement-misled/

Photo from Folkehelseinstitute

The first *Smittestopp* app centralized data and had to delete it after request from the Norwegian Data Protection Authority



FHI stoppar all innsamling av data i Smittestopp

FHI slettar alle data frå appen Smittestopp og stoppar midlertidig innsamlinga av data etter varsel om forbod frå Datatilsynet.

Hans Ivar Moss Kolseth
Journalist

Vilde Gjerde Lied
Journalist

Mette Kristensen
Journalist

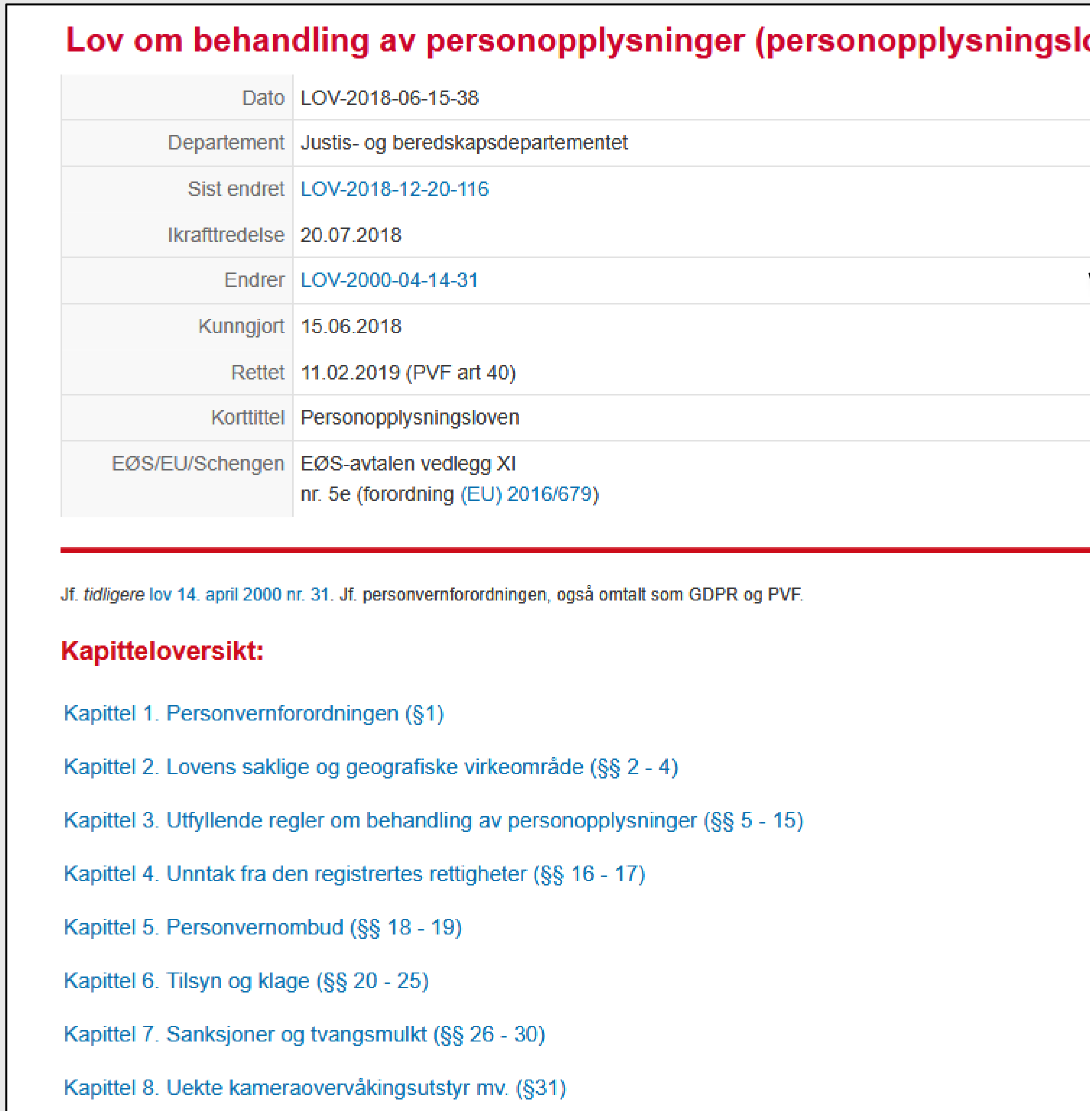
Ugo Fermariello
Journalist

Publisert 15. juni 2020 kl. 06:52
Oppdatert 26. okt. 2020 kl. 19:32

Source:

https://www.nrk.no/norge/fhi-stoppar-all-innsamling-av-data-i-smittestopp-1.15053310

We will examine the rules around collecting data for research

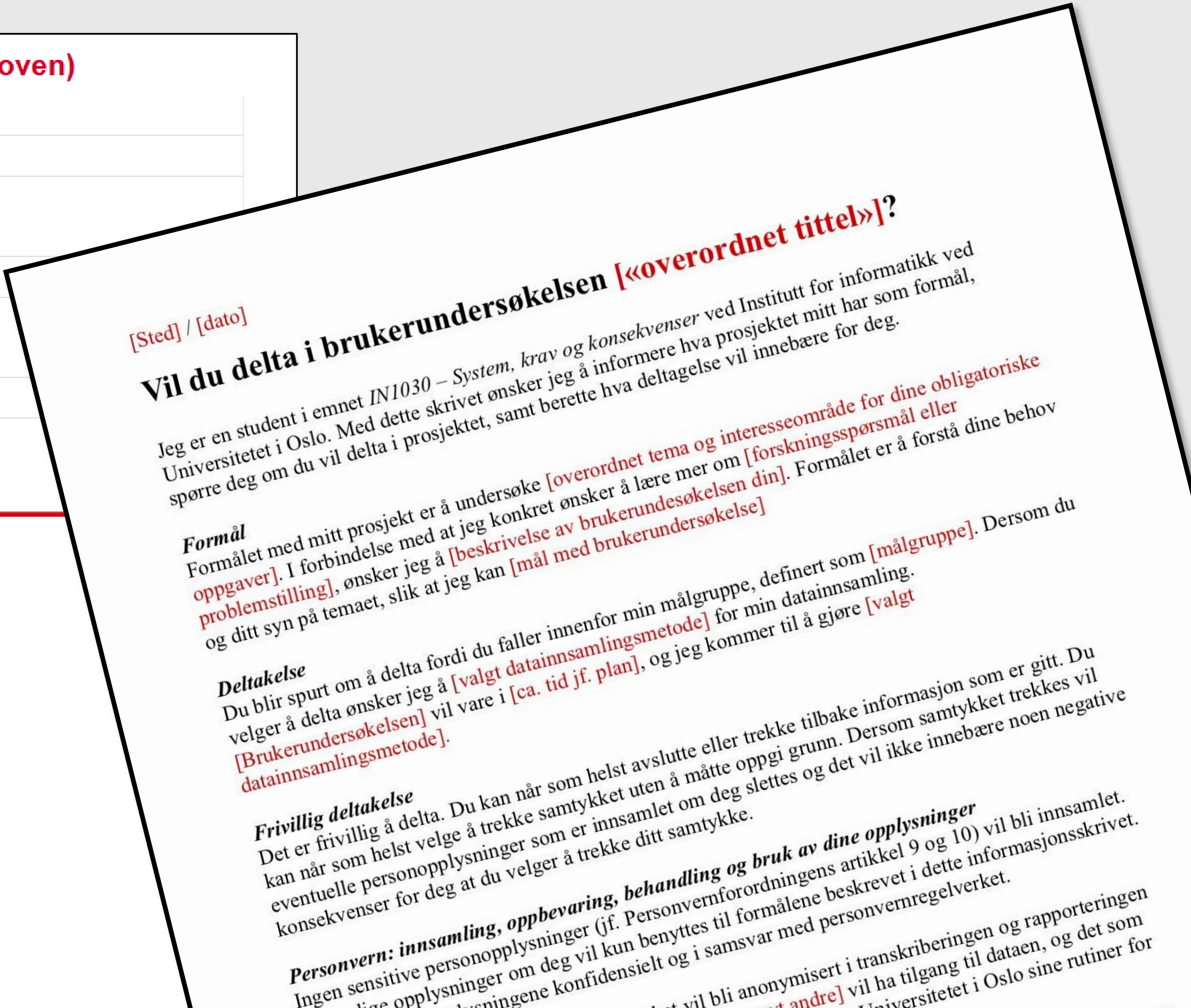


Lov om behandling av personopplysninger (personopplysningsloven)

Dato	LOV-2018-06-15-38
Departement	Justis- og beredskapsdepartementet
Sist endret	LOV-2018-12-20-116
Ikrafttredelse	20.07.2018
Endrer	LOV-2000-04-14-31
Kunngjort	15.06.2018
Rettet	11.02.2019 (PVF art 40)
Korttittel	Personopplysningsloven
EØS/EU/Schengen	EØS-avtalen vedlegg XI nr. 5e (forordning (EU) 2016/679)

Jf. *tidligere* lov 14. april 2000 nr. 31. Jf. personvernforordningen, også omtalt som GDPR og PVF.

Kapitteloversikt:



[Sted] / [dato]

Vil du delta i brukerundersøkelsen [«overordnet tittel»]?

Jeg er en student i emnet *IN1030 – System, krav og konsekvenser* ved Institutt for informatikk ved Universitetet i Oslo. Med dette skrivet ønsker jeg å informere hva prosjektet mitt har som formål, spørre deg om du vil delta i prosjektet, samt berette hva deltagelse vil innebære for deg.

Formål
Formålet med mitt prosjekt er å undersøke [overordnet tema og interesseområde for dine obligatoriske oppgaver]. I forbindelse med at jeg konkret ønsker å lære mer om [forskningsspørsmål eller problemstilling], ønsker jeg å [beskrivelse av brukerundesøkelsen din]. Formålet er å forstå dine behov og ditt syn på temaet, slik at jeg kan [mål med brukerundersøkelse]

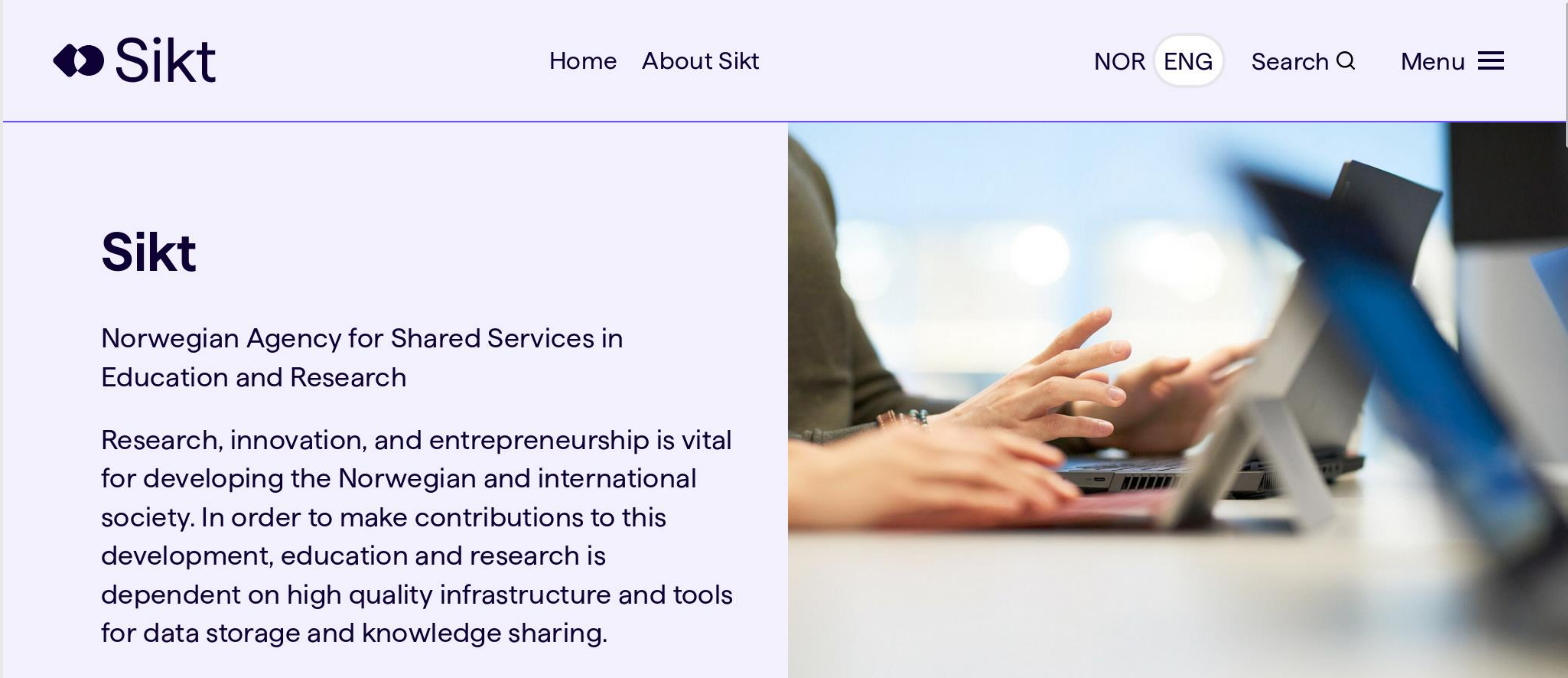
Deltakelse
Du blir spurt om å delta fordi du faller innenfor min målgruppe, definert som [målgruppe]. Dersom du velger å delta ønsker jeg å [valgt datainnsamlingsmetode] for min datainnsamling. [Brukerundersøkelsen] vil vare i [ca. tid jf. plan], og jeg kommer til å gjøre [valgt datainnsamlingsmetode].

Frivillig deltakelse
Det er frivillig å delta. Du kan når som helst avslutte eller trekke tilbake informasjon som er gitt. Du kan når som helst velge å trekke samtykket uten å måtte oppgi grunn. Dersom samtykket trekkes vil eventuelle personopplysninger som er innsamlet om deg slettes og det vil ikke innebære noen negative konsekvenser for deg at du velger å trekke ditt samtykke.

Personvern: innsamling, oppbevaring, behandling og bruk av dine opplysninger
Ingen sensitive personopplysninger (jf. Personvernforordningens artikkel 9 og 10) vil bli innsamlet. lige opplysninger om deg vil kun benyttes til formålene beskrevet i dette informasjonsskrivet. lysningene konfidensielt og i samsvar med personvernregelverket.

et vil bli anonymisert i transkriberingen og rapporteringen t andre] vil ha tilgang til dataen, og det som Universitetet i Oslo sine rutiner for

We will examine what researchers need to do to collect personal information while following the law



Sikt

Norwegian Agency for Shared Services in Education and Research

Research, innovation, and entrepreneurship is vital for developing the Norwegian and international society. In order to make contributions to this development, education and research is dependent on high quality infrastructure and tools for data storage and knowledge sharing.

We will examine some ethics around internet research

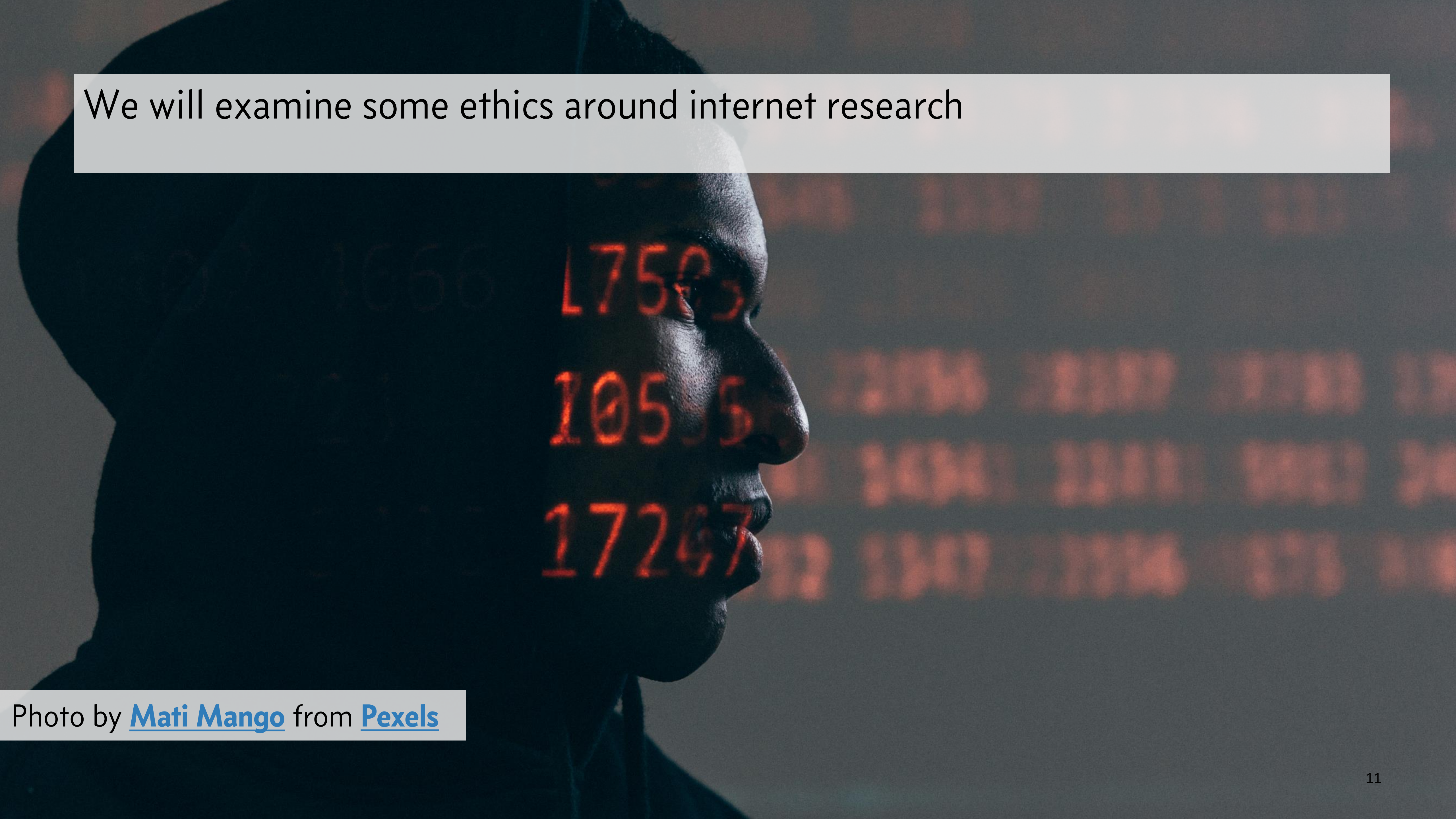


Photo by Mati Mango from Pexels

I am not a lawyer (IANAL), so feel free to examine more

Photo by [Tingey Injury Law Firm] on [Unsplash]

I won't discuss methods for keeping data secure and private

UNIVERSITY OF OSLO

← Research

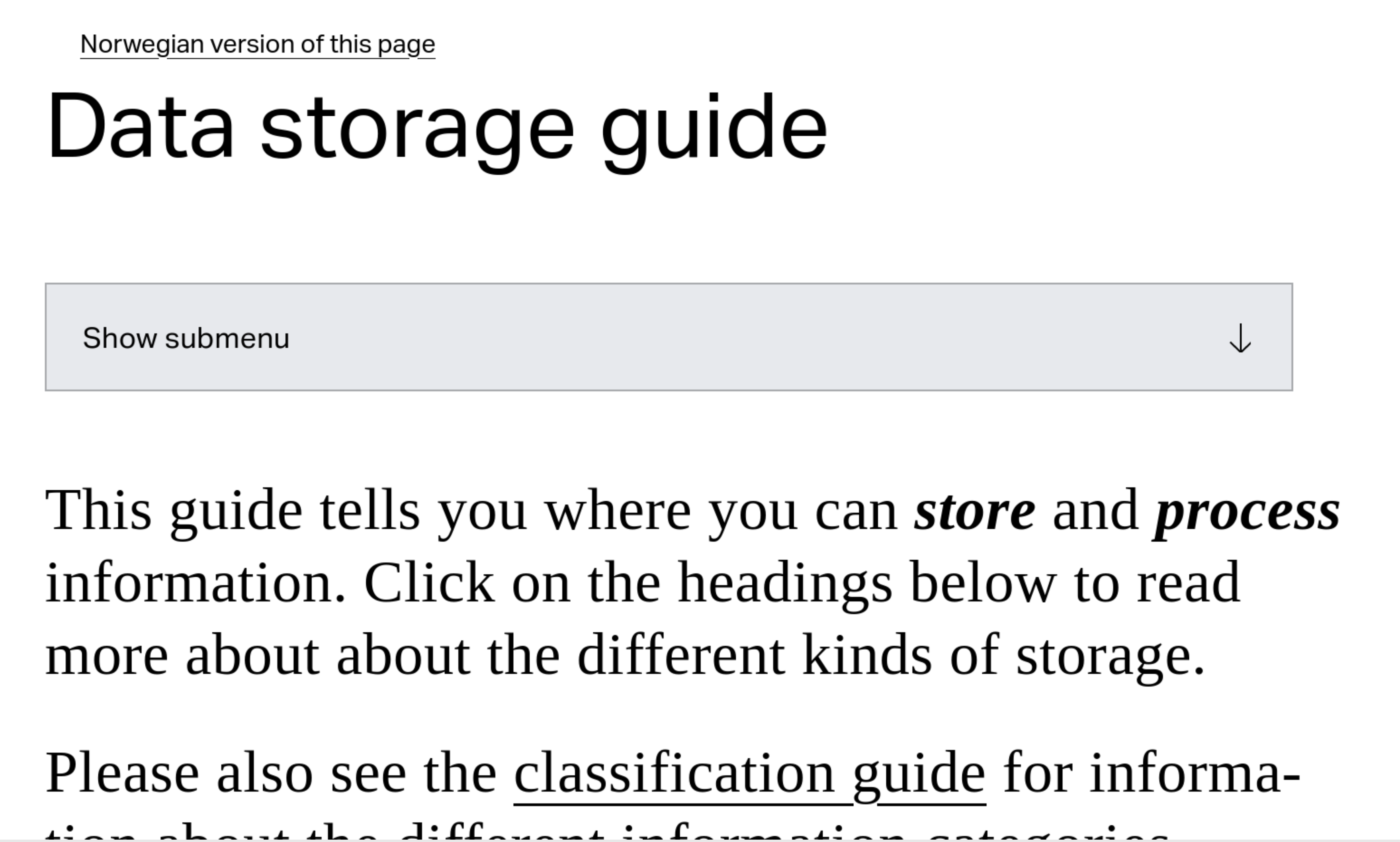
Norwegian version of this page

Services for sensitive data (TSD)

What is TSD?

- provides a platform for researchers at UiO and other public research institutions
- can collect, store and analyze sensitive research data in a secure environment

UiO has a data storage guide that can provide additional information about what data can be stored where



Norwegian version of this page

Data storage guide

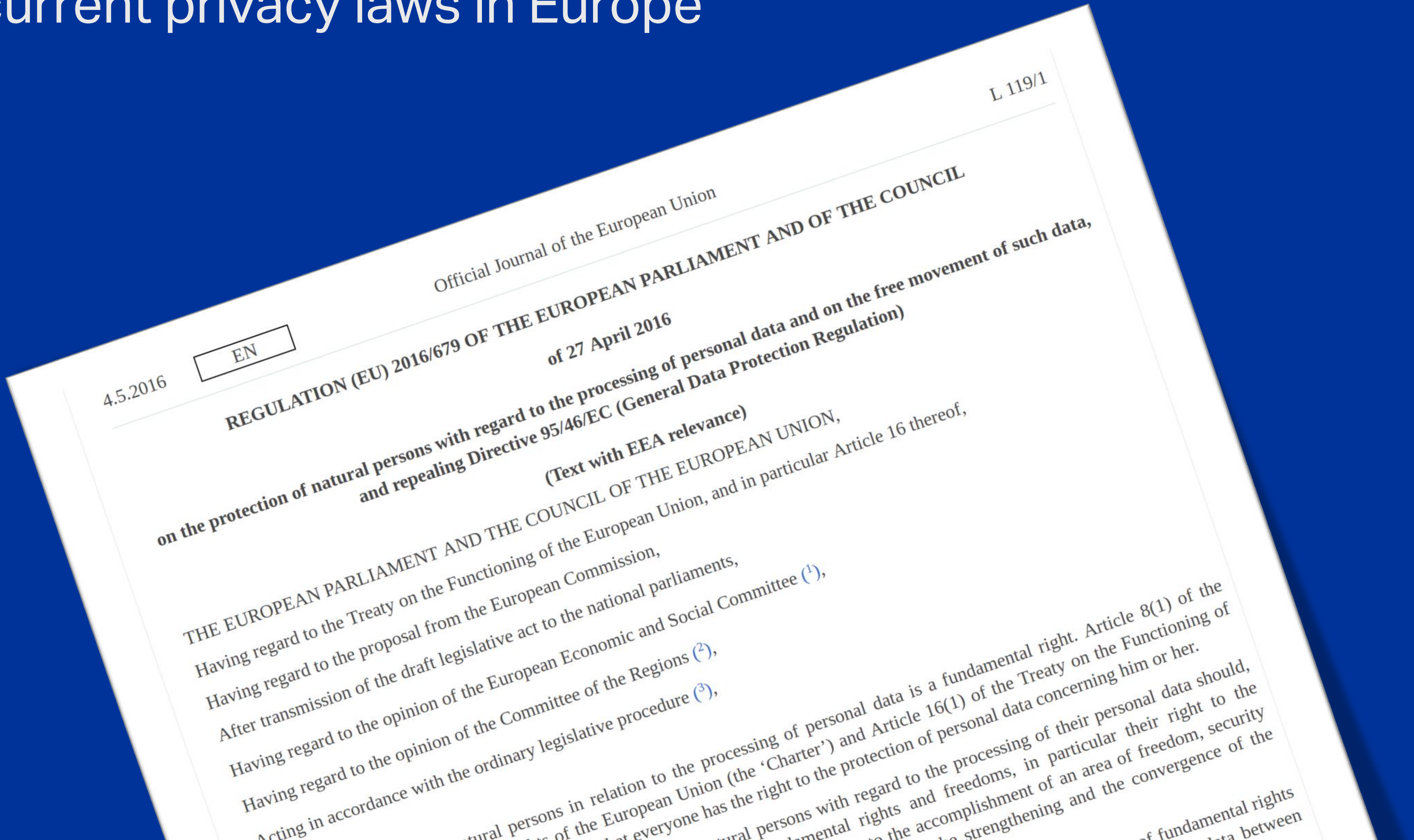
This guide tells you where you can ***store*** and ***process*** information. Click on the headings below to read more about about the different kinds of storage.

Please also see the classification guide for information about the different information categories

Research that involves people has several guidelines to regulate processing of data

- Norwegian personal data act (2018)—Includes GDPR
- European Convention on Human Rights
- UN Declaration of Human Rights
- The Belmont Report
- The Declaration of Helsinki
- The Nuremberg Code

EUs General Data Protection Regulation (GDPR) forms the basis for our current privacy laws in Europe



4.5.2016 EN Official Journal of the European Union L 119/1

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee [1],

Having regard to the opinion of the Committee of the Regions [2],

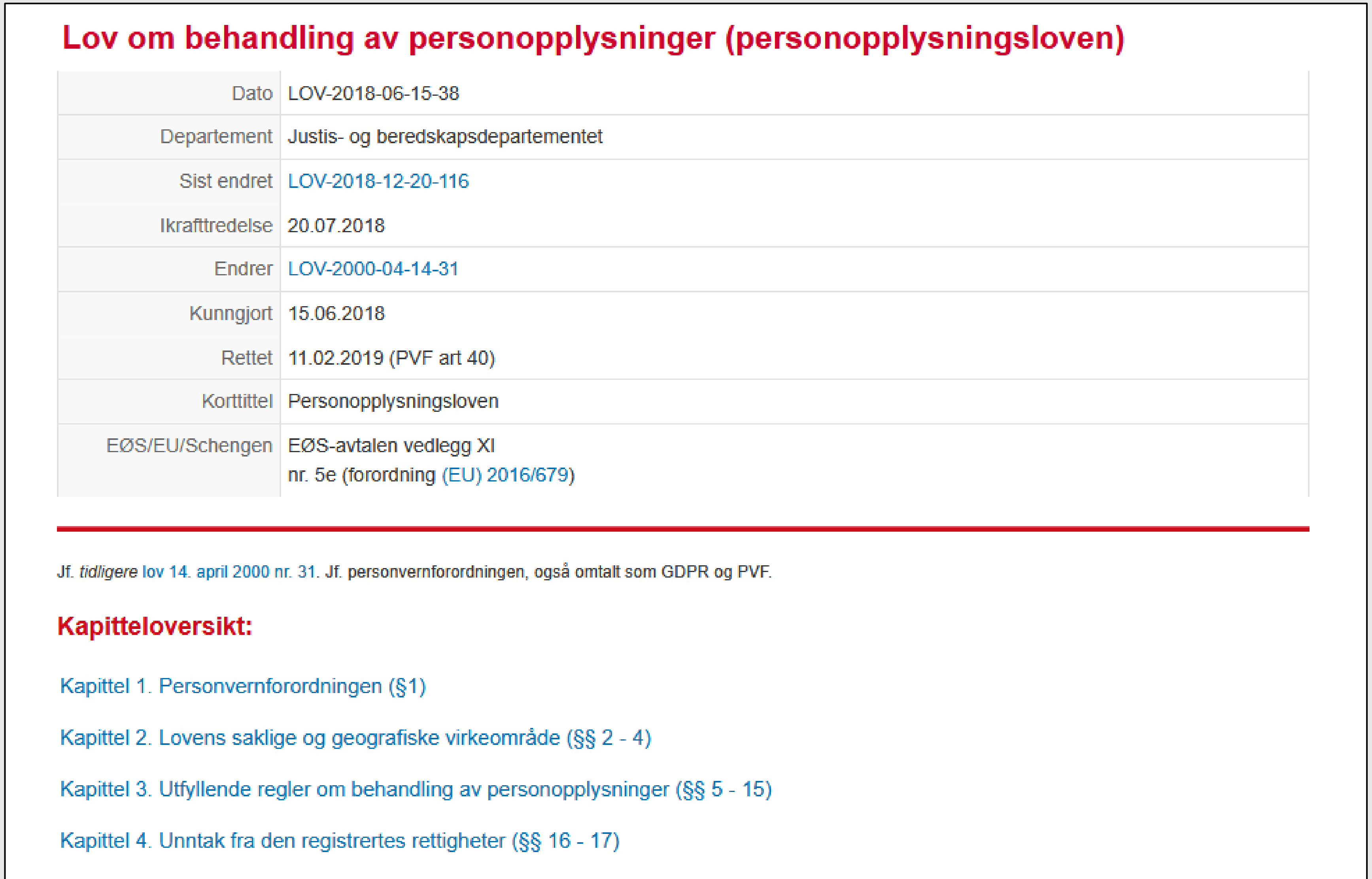
Acting in accordance with the ordinary legislative procedure [3],

tural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the ts of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of at everyone has the right to the protection of personal data concerning him or her.

tural persons with regard to the processing of their personal data should, amental rights and freedoms, in particular their right to the o the accomplishment of an area of freedom, security ke strengthening and the convergence of the

f fundamental rights ata between

Norway implements GDPR through the personal data law *personopplysningsloven* (LOV-2018-06-15-38)



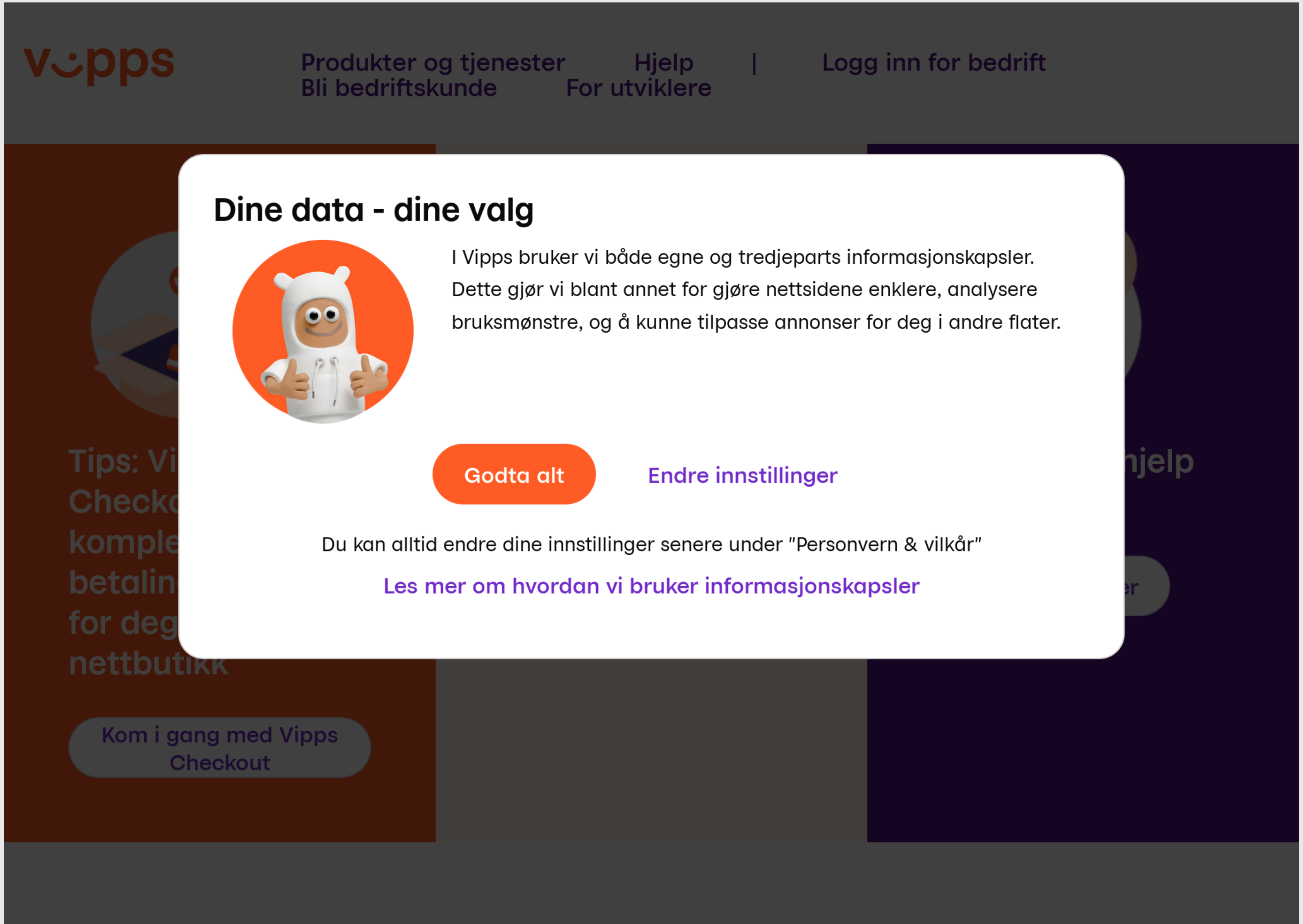
Lov om behandling av personopplysninger (personopplysningsloven)

Dato	LOV-2018-06-15-38
Departement	Justis- og beredskapsdepartementet
Sist endret	LOV-2018-12-20-116
Ikrafttredelse	20.07.2018
Endrer	LOV-2000-04-14-31
Kunngjort	15.06.2018
Rettet	11.02.2019 (PVF art 40)
Korttittel	Personopplysningsloven
EØS/EU/Schengen	EØS-avtalen vedlegg XI nr. 5e (forordning (EU) 2016/679)

Jf. *tidligere* lov 14. april 2000 nr. 31. Jf. personvernforordningen, også omtalt som GDPR og PVF.

Kapitteloversikt:

We experience consequences of the GDPR when we use the internet



Produkter og tjenester
Hjelp
|
Logg inn for bedrift
Bli bedriftskunde
For utviklere

Dine data - dine valg

I Vipps bruker vi både egne og tredjeparts informasjonskapsler. Dette gjør vi blant annet for gjøre nettsidene enklere, analysere bruksmønstre, og å kunne tilpasse annonser for deg i andre flater.

Godta alt

Endre innstillinger

Du kan alltid endre dine innstillinger senere under "Personvern & vilkår"

Les mer om hvordan vi bruker informasjonskapsler

Tips: Vi
Checko
komple
betalin
for deg
nettbutikk

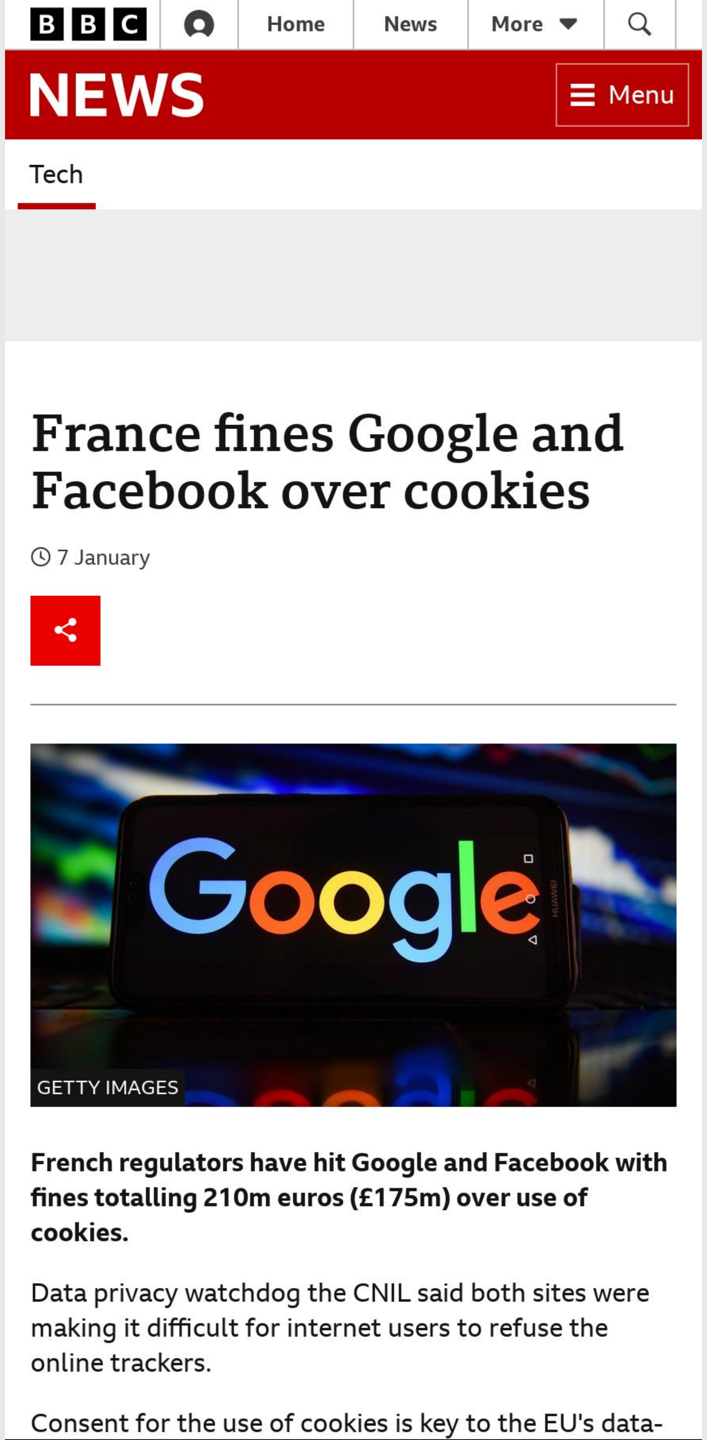
Kom i gang med Vipps Checkout

hjelp

er

Companies also get to experience the GDPR

Source: BBC
7 January 2022
https://www.bbc.com/news/technology-59909647



France fines Google and Facebook over cookies

7 January

GETTY IMAGES

French regulators have hit Google and Facebook with fines totalling 210m euros (£175m) over use of cookies.

Data privacy watchdog the CNIL said both sites were making it difficult for internet users to refuse the online trackers.

Consent for the use of cookies is key to the EU's data-

There are several terms one should know when working with data about people

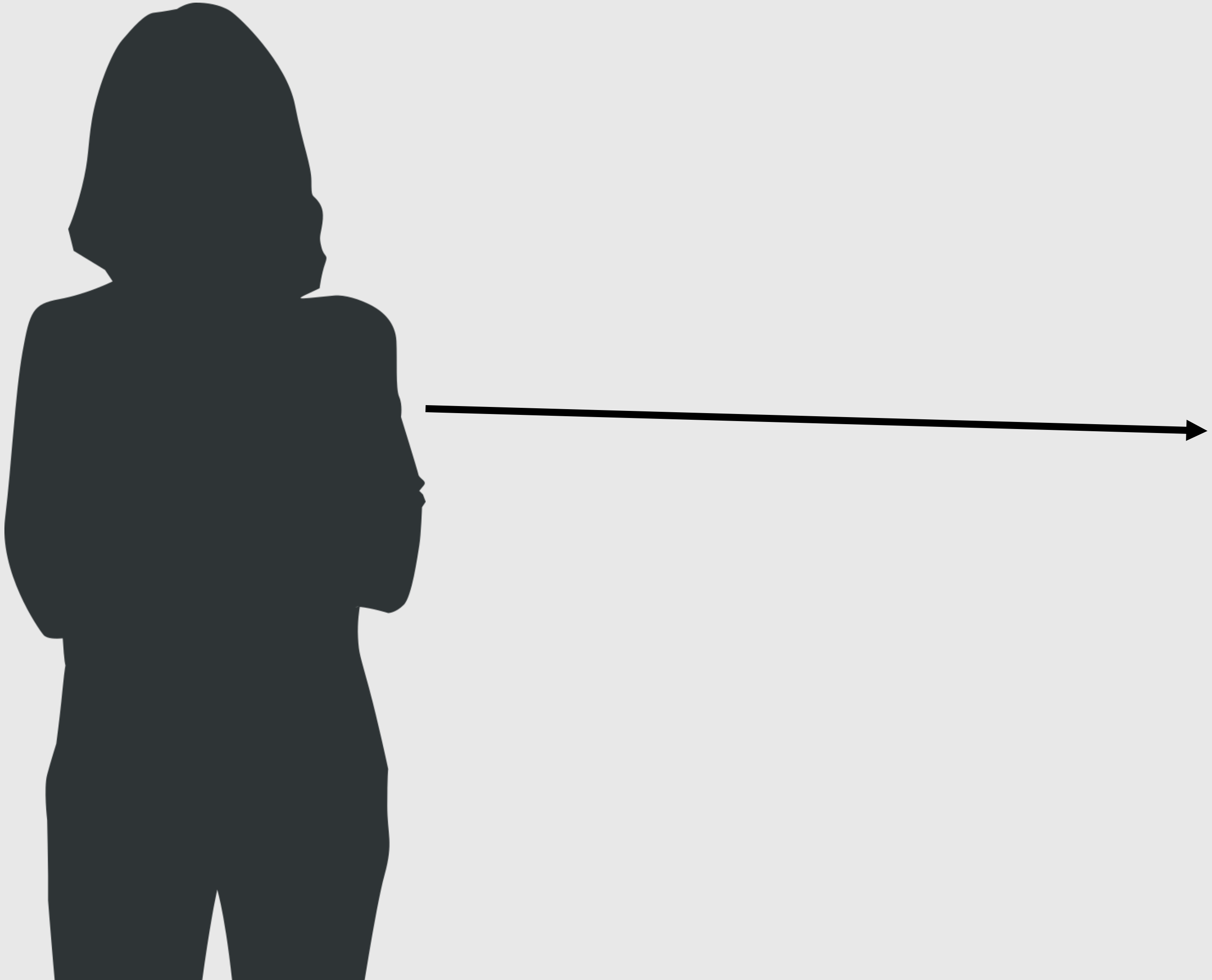
- Data subject
- Personal Data
- Processing
- Controller
- Processor

Data subject (*den registrerte*) is a natural person

Personal data (*personopplysninger*) is any information relating to an identified or identifiable natural person (either directly or indirectly)

Contents:
Facts
by weight:
% Daily Value*
Also conta

Personal data (*personopplysninger*) is any information relating to an identified or identifiable natural person (either directly or indirectly)



Kari Nordmann
Fake street 123
Oslo
Norway

Processing (*Behandling*) is any series of operations done to personal data

Photo by Christina Morillo from Pexels

Controller (*behandlingsansvarlig*) is the person who determines the purpose of processing of the personal data and the means

Photo by Christina Morillo from Pexels

Processor (*databehandler*) is the person who processes the personal data on behalf of the controller

Photo by Christina Morillo from Pexels

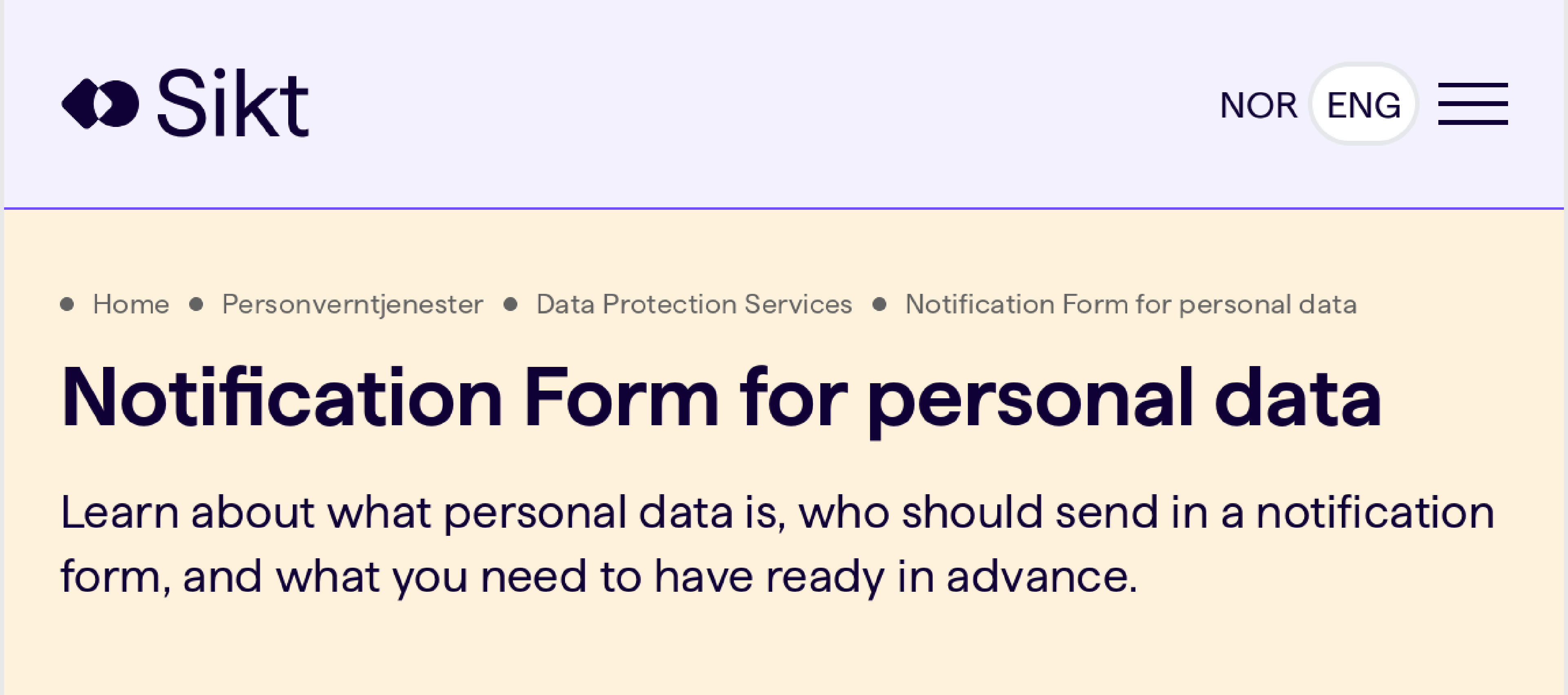
Discussion point

Think about your research projects. Do you think that you will be collecting personal data? Why or why not?

Are there ways that you think can avoid collecting personal data?

GDPR specifies requirements for the controller doing research that involves collecting or processing personal data

Main requirement for researchers: all research involving personal data must be reported using a notification form to the privacy ombudsman for research. For universities in Norway this is Sikt: Norwegian Agency for Shared Services in Education and Research



- Home
- Personverntjenester
- Data Protection Services
- Notification Form for personal data

Notification Form for personal data

Learn about what personal data is, who should send in a notification form, and what you need to have ready in advance.

Basic rules for “Should I contact Sikt?”

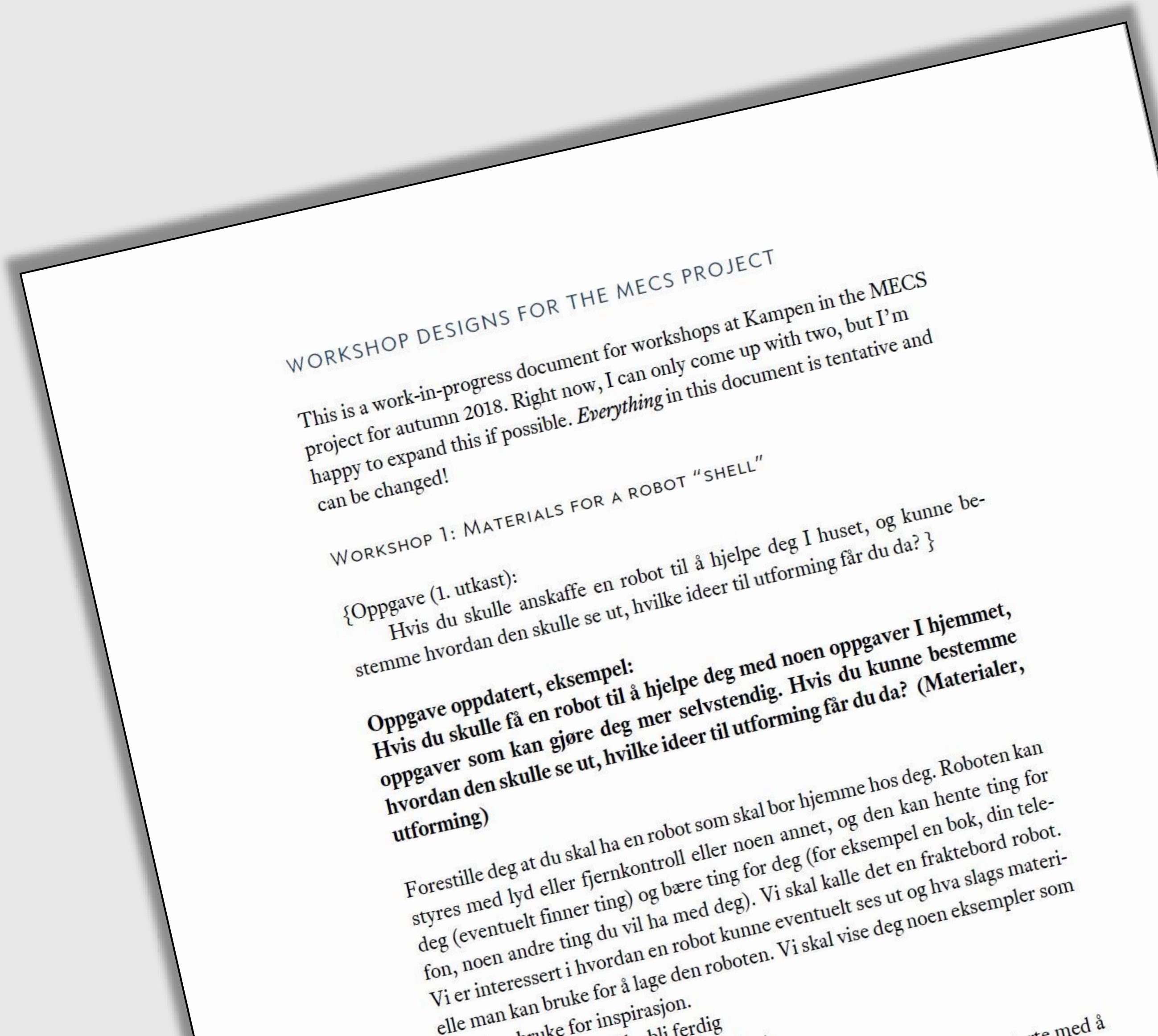
1. Recording or processing of information about individuals by electronic means.
2. A manual register containing special categories of personal data will be created

Filling out the notification form (*meldeskjema*) is straight forward; just remember…

Plan your study carefully: roundtrips with Sikt will take time

Photo by **fauxels** from **Pexels**

Have your paperwork in order; all documents need to be included



WORKSHOP DESIGNS FOR THE MECS PROJECT

This is a work-in-progress document for workshops at Kampen in the MECS project for autumn 2018. Right now, I can only come up with two, but I'm happy to expand this if possible. *Everything* in this document is tentative and can be changed!

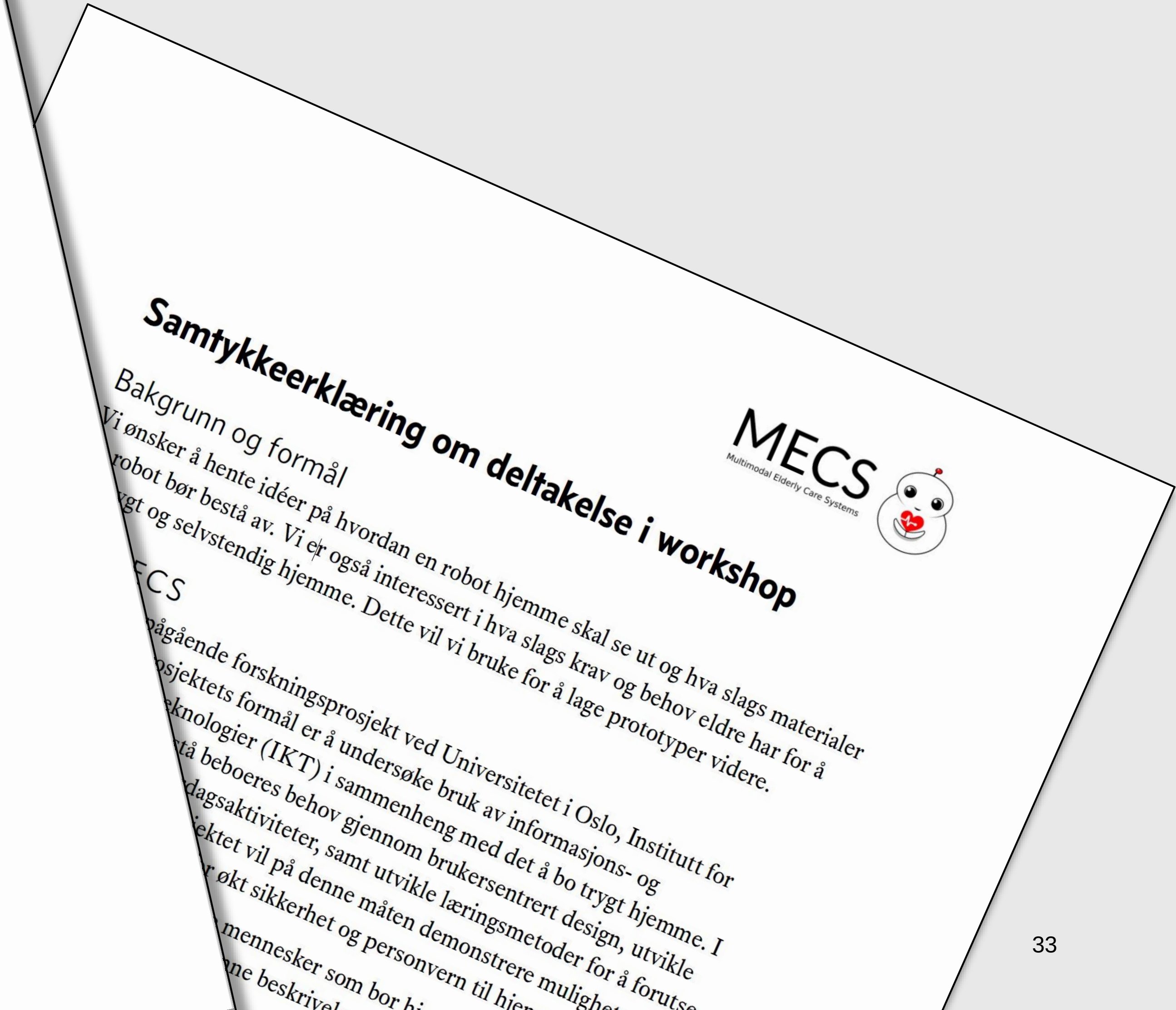
WORKSHOP 1: MATERIALS FOR A ROBOT "SHELL"

{Oppgave (1. utkast):

Hvis du skulle anskaffe en robot til å hjelpe deg I huset, og kunne bestemme hvordan den skulle se ut, hvilke ideer til utforming får du da? }

Oppgave oppdatert, eksempel:
Hvis du skulle få en robot til å hjelpe deg med noen oppgaver I hjemmet, oppgaver som kan gjøre deg mer selvstendig. Hvis du kunne bestemme hvordan den skulle se ut, hvilke ideer til utforming får du da? (Materialer, utforming)

Forestille deg at du skal ha en robot som skal bor hjemme hos deg. Roboten kan styres med lyd eller fjernkontroll eller noen annet, og den kan hente ting for deg (eventuelt finner ting) og bære ting for deg (for eksempel en bok, din telefon, noen andre ting du vil ha med deg). Vi skal kalle det en fraktebord robot. Vi er interessert i hvordan en robot kunne eventuelt ses ut og hva slags materielle man kan bruke for å lage den roboten. Vi skal vise deg noen eksempler som
bruke for inspirasjon.
bli ferdig
te med å



Samtykkeerklæring om deltakelse i workshop

MECS
Multimodal Elderly Care Systems

Bakgrunn og formål

Vi ønsker å hente idéer på hvordan en robot hjemme skal se ut og hva slags materialer
robot bør bestå av. Vi er også interessert i hva slags krav og behov eldre har for å
ygt og selvstendig hjemme. Dette vil vi bruke for å lage prototyper videre.

ECS

pågående forskningsprosjekt ved Universitetet i Oslo, Institutt for
osjektets formål er å undersøke bruk av informasjons- og
eknologier (IKT) i sammenheng med det å bo trygt hjemme. I
stå beboeres behov gjennom brukersentrert design, utvikle
dagsaktiviteter, samt utvikle læringsmetoder for å forutse
iektet vil på denne måten demonstrere muligheter
r økt sikkerhet og personvern til hjem
mennesker som bor hj
ne beskrivel

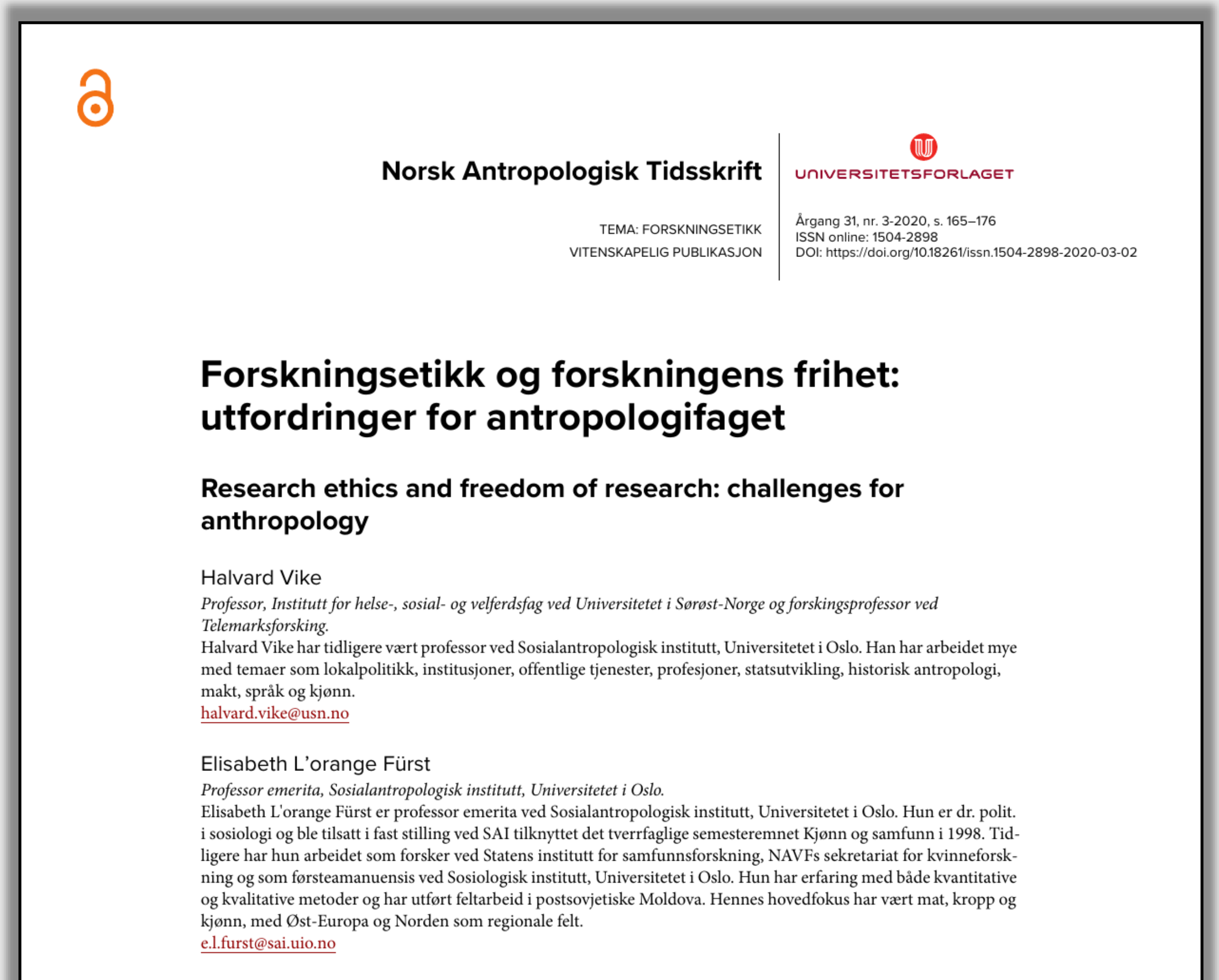
Processing the form takes approximately 30 days: **no data collection** until the form is processed

Photo by Anete Lusina from Pexels

When in doubt, talk with your advisor or contact Sikt

Sikt is there to help you comply with the law; there are not there to play “gotcha”.

Vike and L’orange Fürst pointed out several issues with the GDPR, NSD (now Sikt), and data collection for anthropologists



Norsk Antropologisk Tidsskrift

UNIVERSITETSFORLAGET

TEMA: FORSKNINGSETIKK
VITENSKAPELIG PUBLIKASJON

Årgang 31, nr. 3-2020, s. 165–176
ISSN online: 1504-2898
DOI: https://doi.org/10.18261/issn.1504-2898-2020-03-02

Forskningsetikk og forskningens frihet: utfordringer for antropologifaget

Research ethics and freedom of research: challenges for anthropology

Halvard Vike

Professor, Institutt for helse-, sosial- og velferdsfag ved Universitetet i Sørøst-Norge og forskingsprofessor ved Telemarksforsking.

Halvard Vike har tidligere vært professor ved Sosialantropologisk institutt, Universitetet i Oslo. Han har arbeidet mye med temaer som lokalpolitikk, institusjoner, offentlige tjenester, profesjoner, statsutvikling, historisk antropologi, makt, språk og kjønn.

halvard.vike@usn.no

Elisabeth L’orange Fürst

Professor emerita, Sosialantropologisk institutt, Universitetet i Oslo.

Elisabeth L'orange Fürst er professor emerita ved Sosialantropologisk institutt, Universitetet i Oslo. Hun er dr. polit. i sosiologi og ble tilsatt i fast stilling ved SAI tilknyttet det tverrfaglige semesteremnet Kjønn og samfunn i 1998. Tidligere har hun arbeidet som forsker ved Statens institutt for samfunnsforskning, NAVFs sekretariat for kvinneforskning og som førsteamanuensis ved Sosiologisk institutt, Universitetet i Oslo. Hun har erfaring med både kvantitative og kvalitative metoder og har utført feltarbeid i postsovjetiske Moldova. Hennes hovedfokus har vært mat, kropp og kjønn, med Øst-Europa og Norden som regionale felt.

e.l.furst@sai.uio.no

Source:
Vike, H., & L’orange Fürst, E. (2020). Forskningsetikk og forskningens frihet: Utfordringerfor antropologifaget. *Norsk antropologisk tidsskrift*, *31*(3), 165–176. https://doi.org/10.18261/issn.1504-2898-2020-03-02

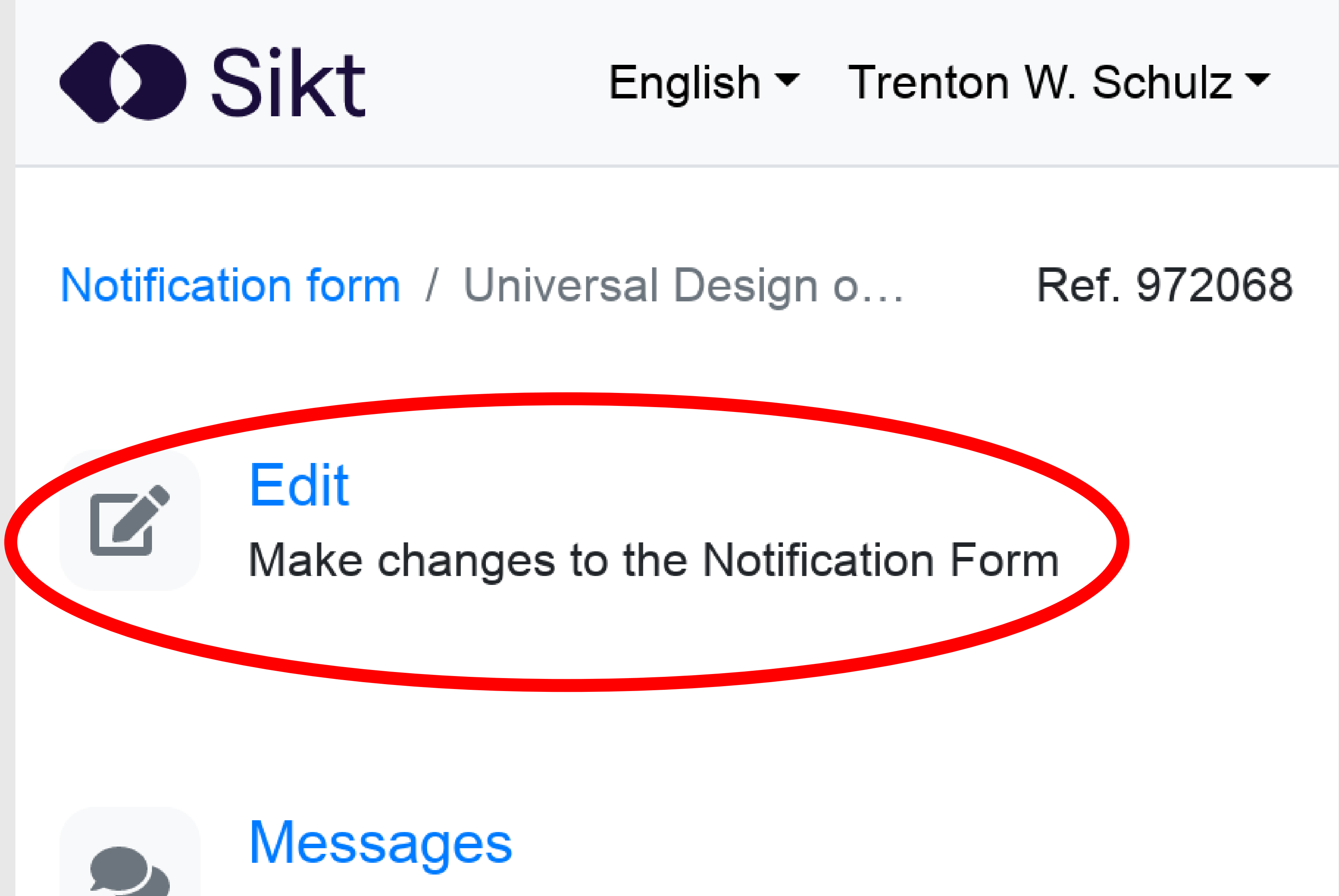
New since December 2022, Sikt offers automatic evaluation of some notification forms

The form is evaluated by a machine with no advisor looking at the form. This is done for forms that collect personal information that has low risk of causing issues. There are multiple criteria, but the big points are:

- The project does not collect special categories of personal data.
- The length of the project is limited
- The number of participants is low
- All participants get individual information
- All participants are over 15-years old

Informed consent forms and other attachments are **not** examined, you are strongly encouraged to use templates from Sikt to be compliant.

If you research changes and you need to collect different data, you need to inform about changes



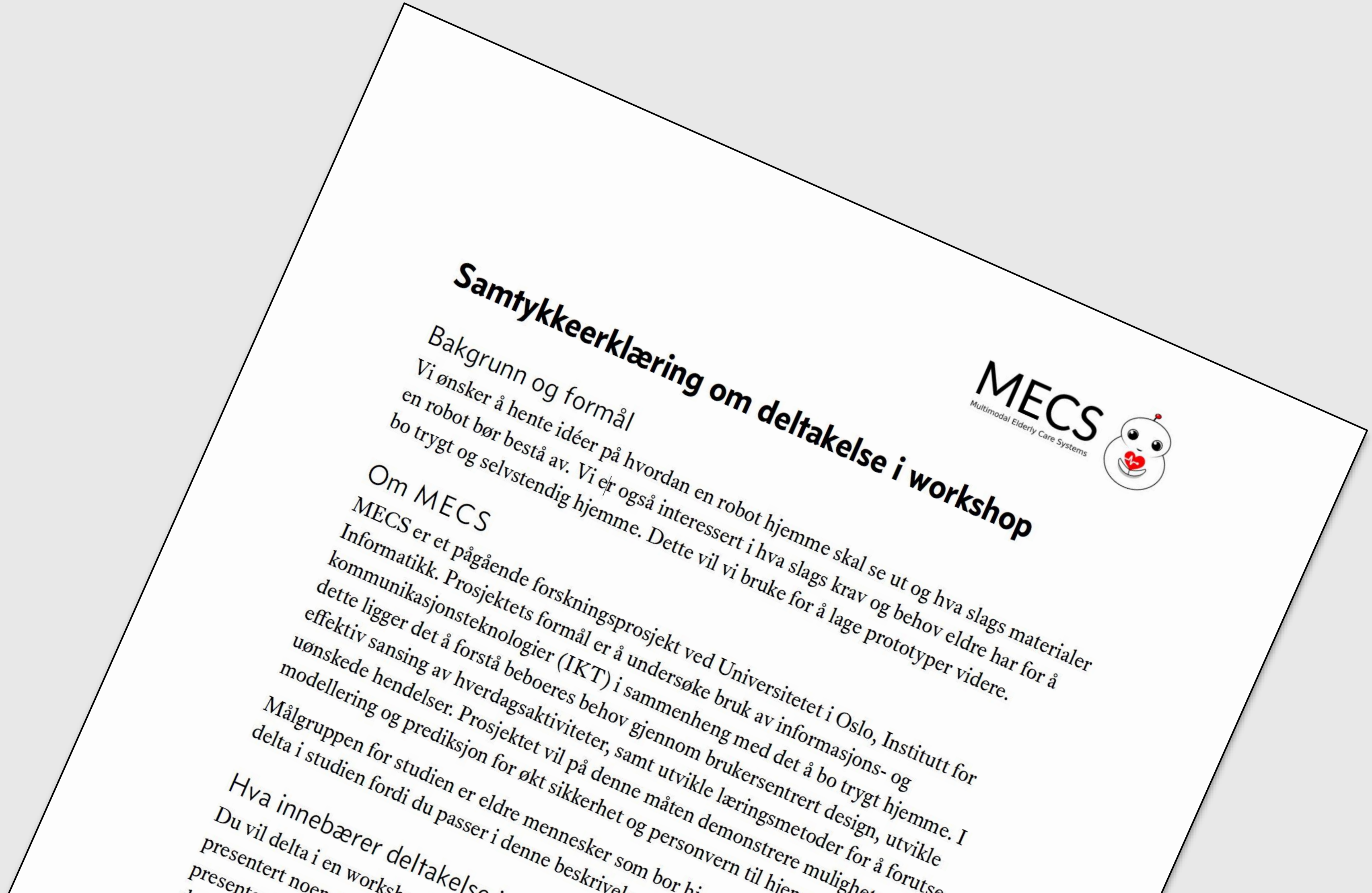
Notification form / Universal Design o… Ref. 972068

Edit

Make changes to the Notification Form

Messages

You cannot reuse your data for another purpose; you must ask for new consent



MECS
Multimodal Elderly Care Systems

Samtykkeerklæring om deltakelse i workshop

Bakgrunn og formål

Vi ønsker å hente idéer på hvordan en robot hjemme skal se ut og hva slags materialer en robot bør bestå av. Vi er også interessert i hva slags krav og behov eldre har for å bo trygt og selvstendig hjemme. Dette vil vi bruke for å lage prototyper videre.

Om MECS

MECS er et pågående forskningsprosjekt ved Universitetet i Oslo, Institutt for Informatikk. Prosjektets formål er å undersøke bruk av informasjons- og kommunikasjonsteknologier (IKT) i sammenheng med det å bo trygt hjemme. I dette ligger det å forstå beboeres behov gjennom brukersentrert design, utvikle effektiv sansing av hverdagsaktiviteter, samt utvikle læringsmetoder for å forutse uønskede hendelser. Prosjektet vil på denne måten demonstrere muligheter modellering og prediksjon for økt sikkerhet og personvern til hjem

Målgruppen for studien er eldre mennesker som bor hj delta i studien fordi du passer i denne beskrivel

Hva innebærer deltakelse i

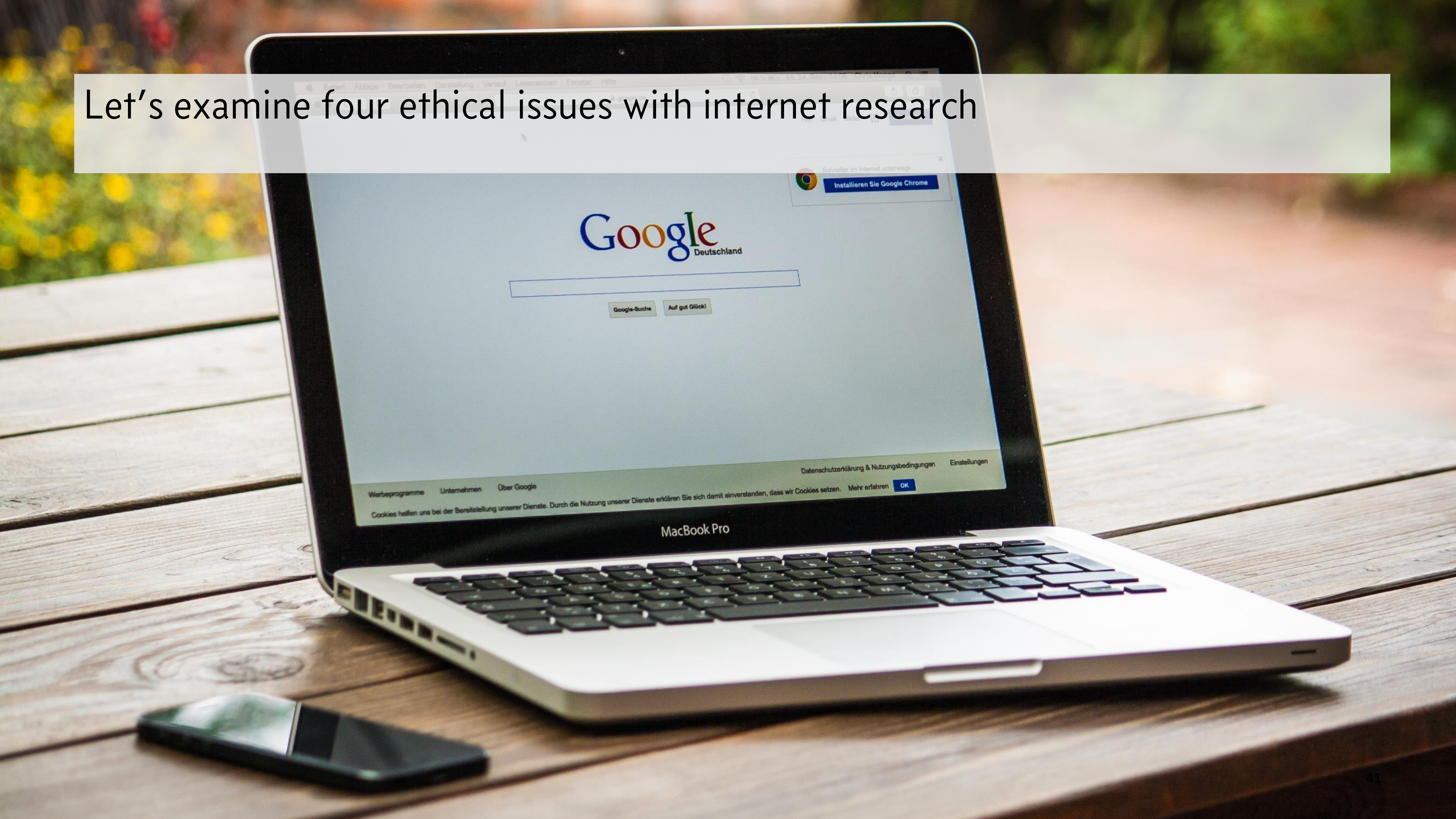
Du vil delta i en worksh presentert noen presente

Unless you are working with health data you *probably do not need* approval from REK

REK (*Regional Komiteer for Medisinsk og Helesfaglig Forskningsetikk*) looks at all research projects that involve medicine, health, or research biobanks.

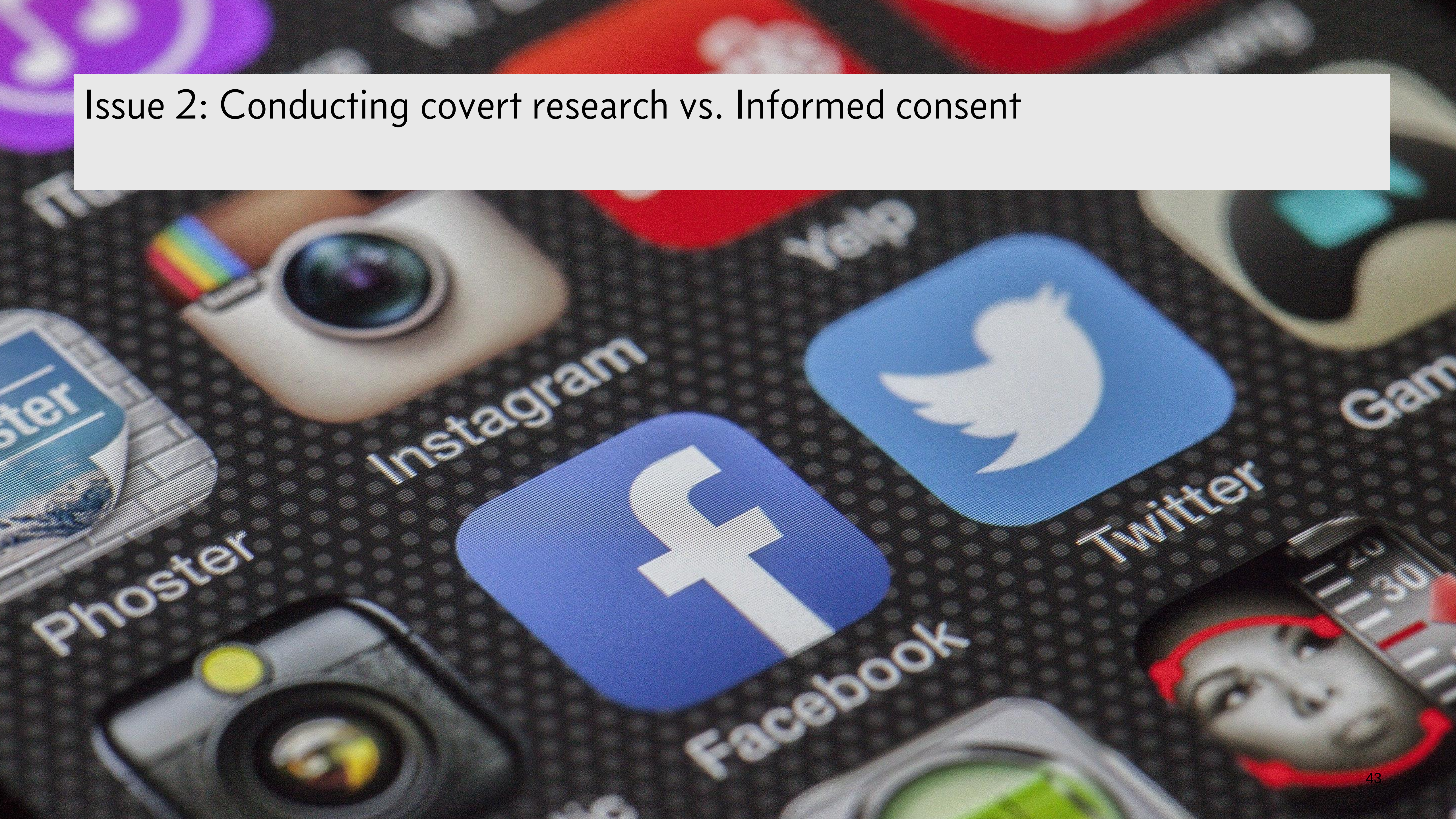
REK’s approval only looks at the health and medical aspects of the research, you still need to fill a form with Sikt if you are processing personal data.

Let’s examine four ethical issues with internet research

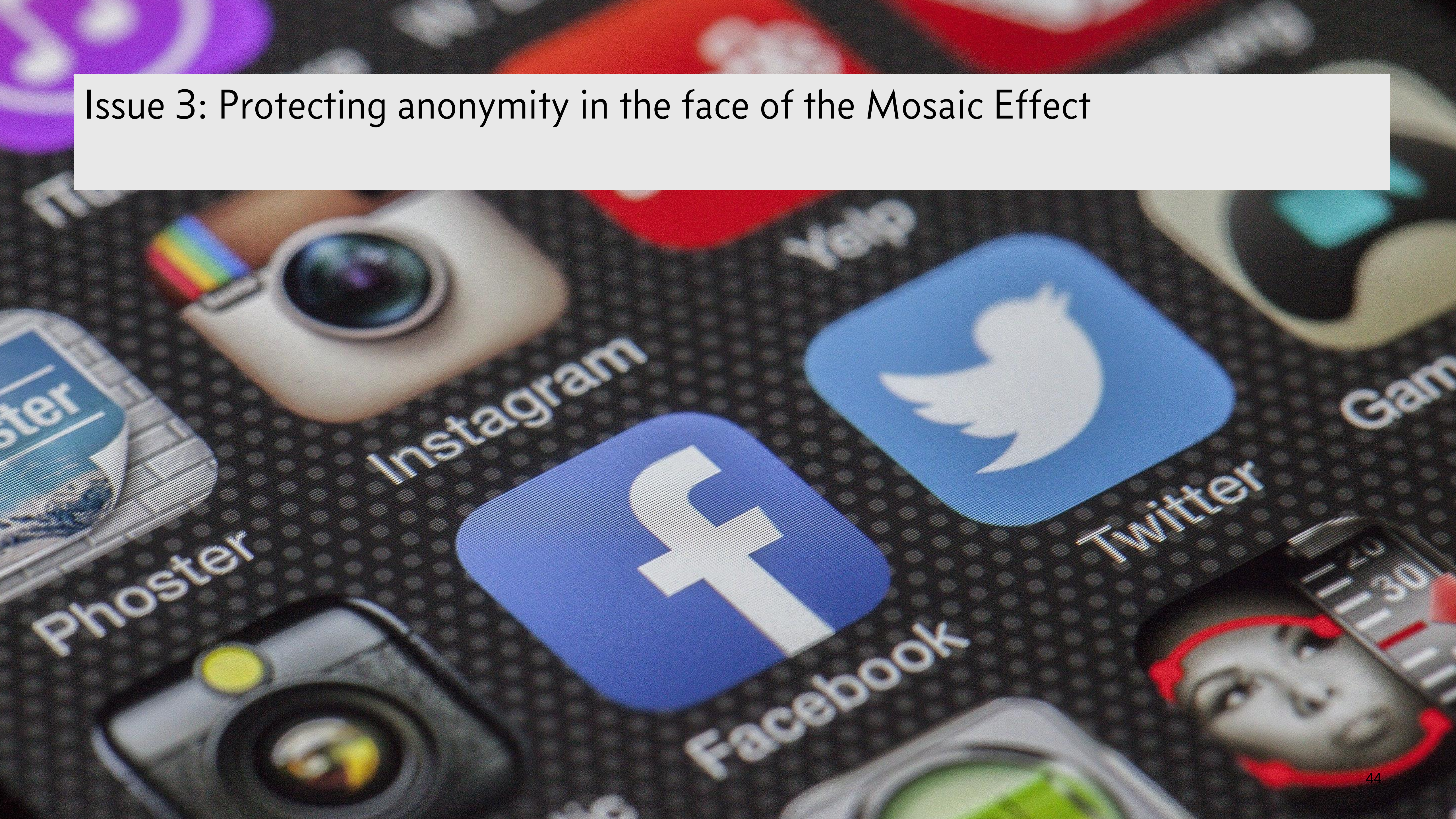

Installieren Sie Google Chrome
Google
Deutschland
Google-Suche
Auf gut Glück!
Datenschutzerklärung & Nutzungsbedingungen
Einstellungen
Werbeprogramme
Unternehmen
Über Google
Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
Mehr erfahren
OK
MacBook Pro

Issue 1: Is online interpersonal media (social media) considered public or private information?

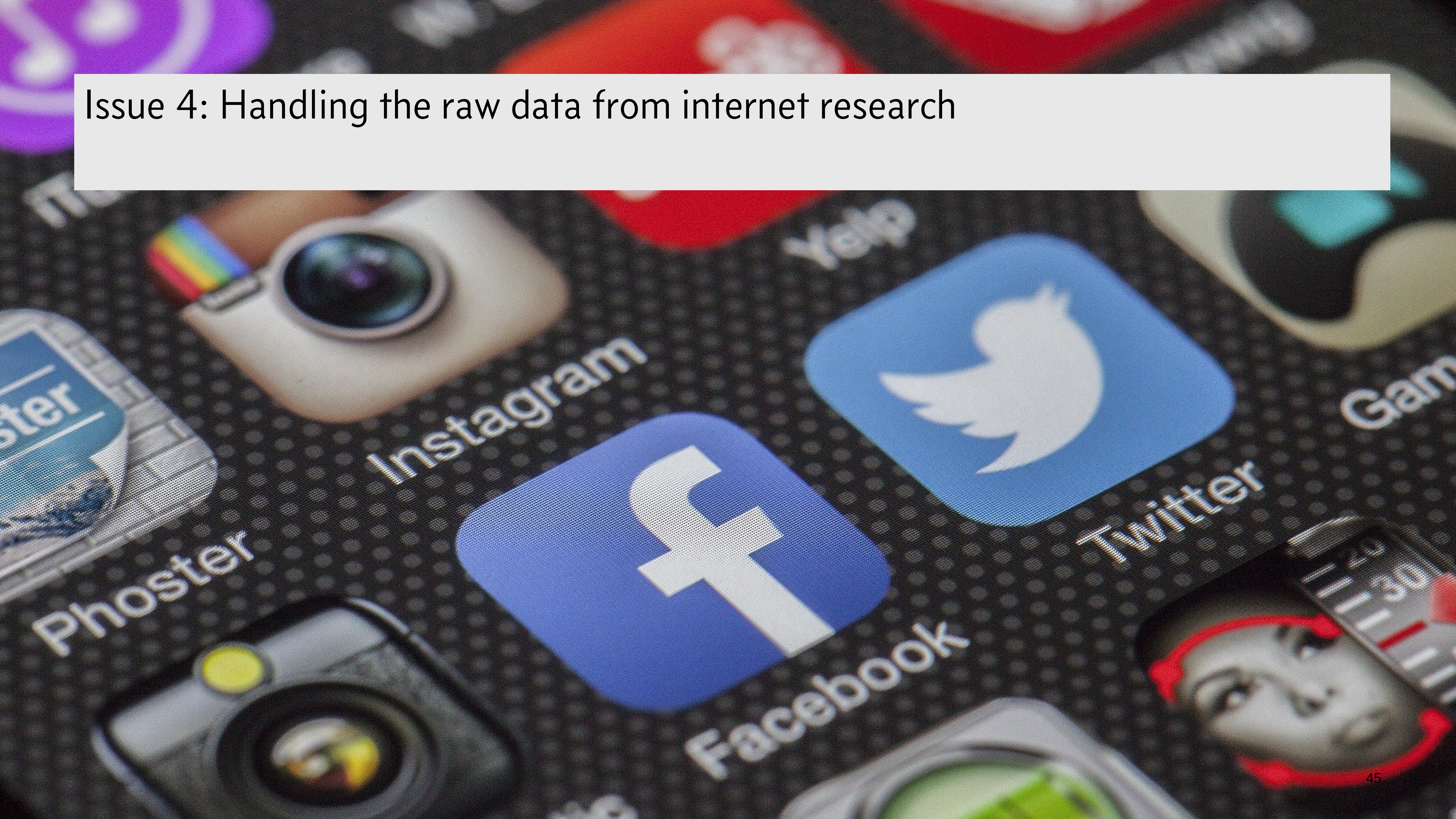
Issue 2: Conducting covert research vs. Informed consent



Issue 3: Protecting anonymity in the face of the Mosaic Effect



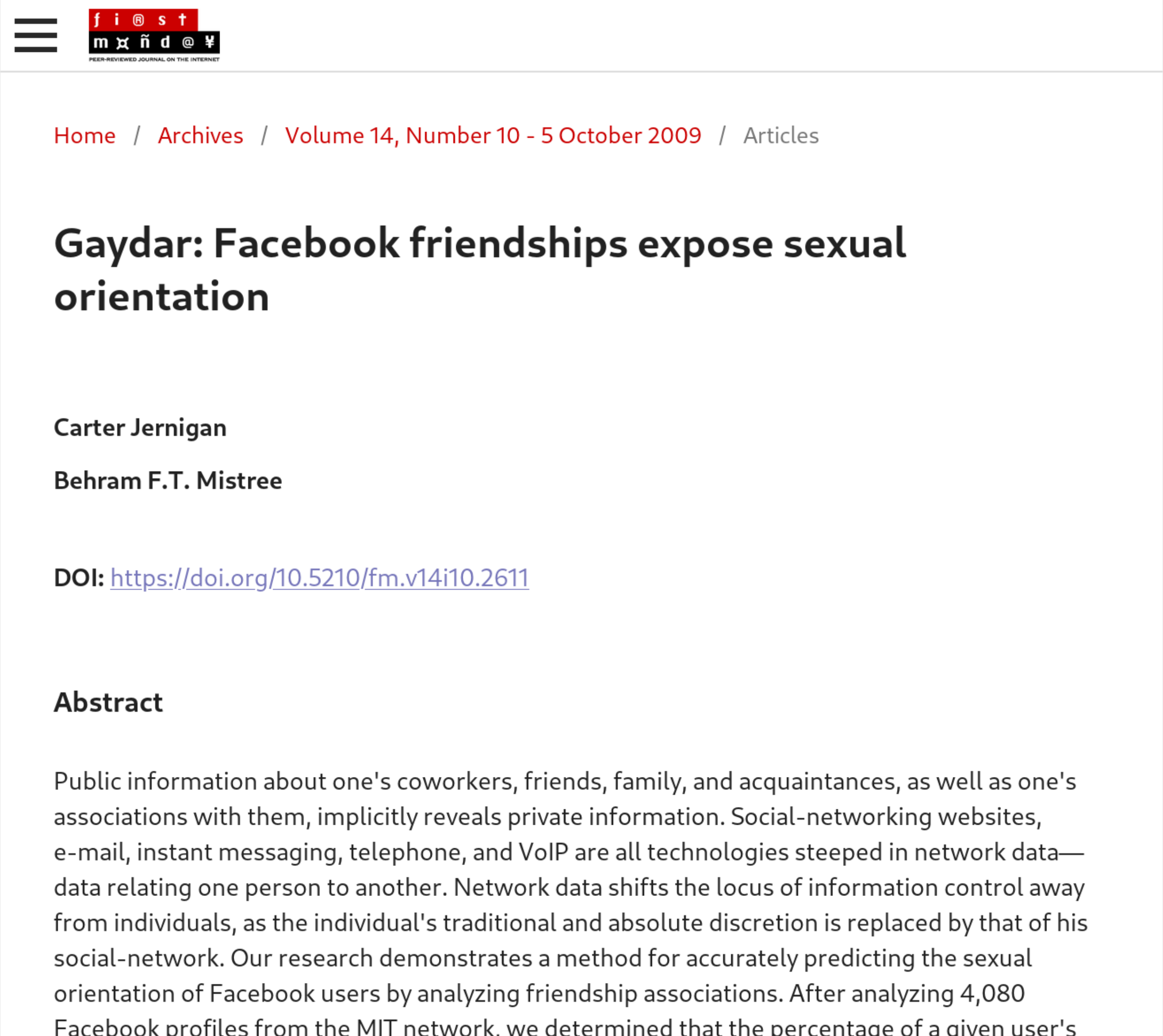
Issue 4: Handling the raw data from internet research



Internet Research Ethic Sources

- Cheltenham and Gloucester *College of Higher Education: Research Ethics: A Handbook of Principles and Procedures.*
- Association of Internet Researchers (AoIR), reports on *Ethical and Legal Aspects of Research on the Internet* http://aoir.org/reports/ethics.pdf
  http://aoir.org/reports/ethics2.pdf
  https://aoir.org/reports/ethics3.pdf

Issue 1: Is social media information private or public + Issue 2 covert research versus informed consent: The *Gaydar* Study



Home / Archives / Volume 14, Number 10 - 5 October 2009 / Articles

Gaydar: Facebook friendships expose sexual orientation

Carter Jernigan

Behram F.T. Mistree

DOI: https://doi.org/10.5210/fm.v14i10.2611

Abstract

Public information about one's coworkers, friends, family, and acquaintances, as well as one's associations with them, implicitly reveals private information. Social-networking websites, e-mail, instant messaging, telephone, and VoIP are all technologies steeped in network data—data relating one person to another. Network data shifts the locus of information control away from individuals, as the individual's traditional and absolute discretion is replaced by that of his social-network. Our research demonstrates a method for accurately predicting the sexual orientation of Facebook users by analyzing friendship associations. After analyzing 4,080 Facebook profiles from the MIT network, we determined that the percentage of a given user's

Source:

Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, *14*(10). https://doi.org/10.5210/fm.v14i10.2611

Jernigan & Mistree's look at the relations between people on Facebook to find a person's orientation; from the abstract

Public information about one's coworkers, friends, family, and acquaintances, as well as one's associations with them, implicitly reveals private information… After analyzing 4,080 Facebook profiles from the MIT network, we determined that the percentage of a given user's friends who self-identify as gay male is strongly correlated with the sexual orientation of that user, and we developed a logistic regression classifier with strong predictive power.

Jernigan & Mistree collected the information without contacting the user; later in the article, emphasis added

> “Our analysis demonstrates a method of classifying sexual orientation of individuals on Facebook, *regardless of whether they chose to disclose that information*. *Facebook users who did not disclose their sexual orientation in their profiles would presumably consider the present research an invasion of privacy*. *Yet this research uses nothing more than information already publicly provided on Facebook; no interaction with subjects was required*. Although we based our research solely on public information, only a limited subset of our results, which contain no personally identifiable information, is presented in this paper to maintain subject confidentiality.”

Discussion point

What are your opinions about how the data was collected?

Do you think it was necessary to gather informed consent?

Does the GDPR provide any sort of guidance in this issue?

AOIR has suggestions about when informed consent is *not necessary*

- Data is collected from the public sphere with no intervention from the persons whose activities are observed and recorded
- The collection of data does not include personal identifiers which, if released, could result in reputational or financial harm to the person whose activities are observed

Issue 3: Protecting anonymity: Researchers have a duty to protect the anonymity of the people in the data

Researchers must take care where the alteration of contexts may reveal the identity of data sets hitherto protected. Particular care should be taken with data that arises from covert … research methods …

– *Research Ethics Handbook*

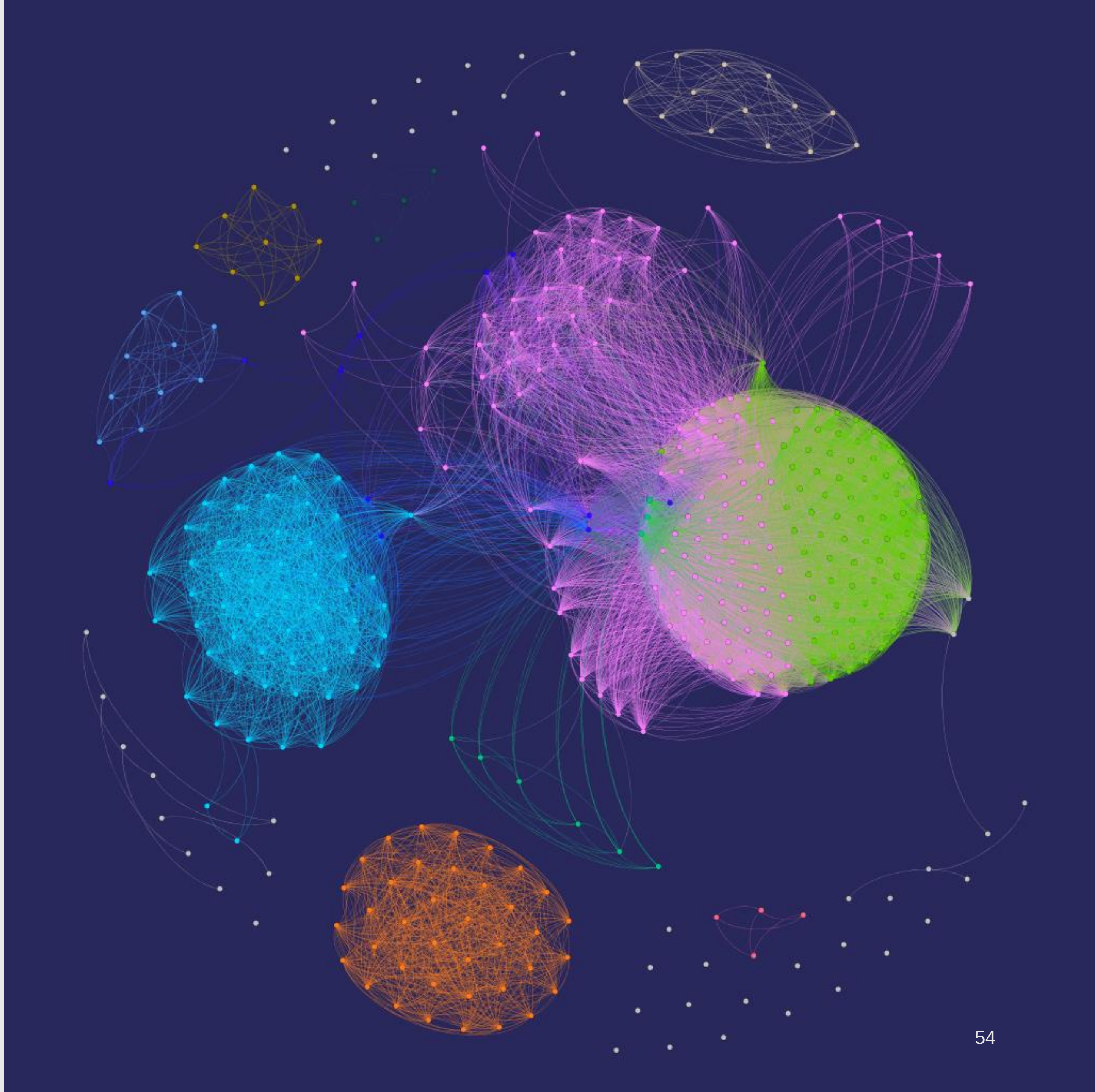
The Mosaic Effect: your anonymous data may reveal more information when combined with other information

> “The Mosaic Effect occurs when the information in an individual dataset, in isolation, may not pose a risk of identifying an individual (or threatening some other important interest such as security), but when combined with other available information, could pose such risk. Before disclosing potential PII [personally identifiable information] or other potentially sensitive information, agencies must consider other publicly available data in any medium and from any source to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern.”

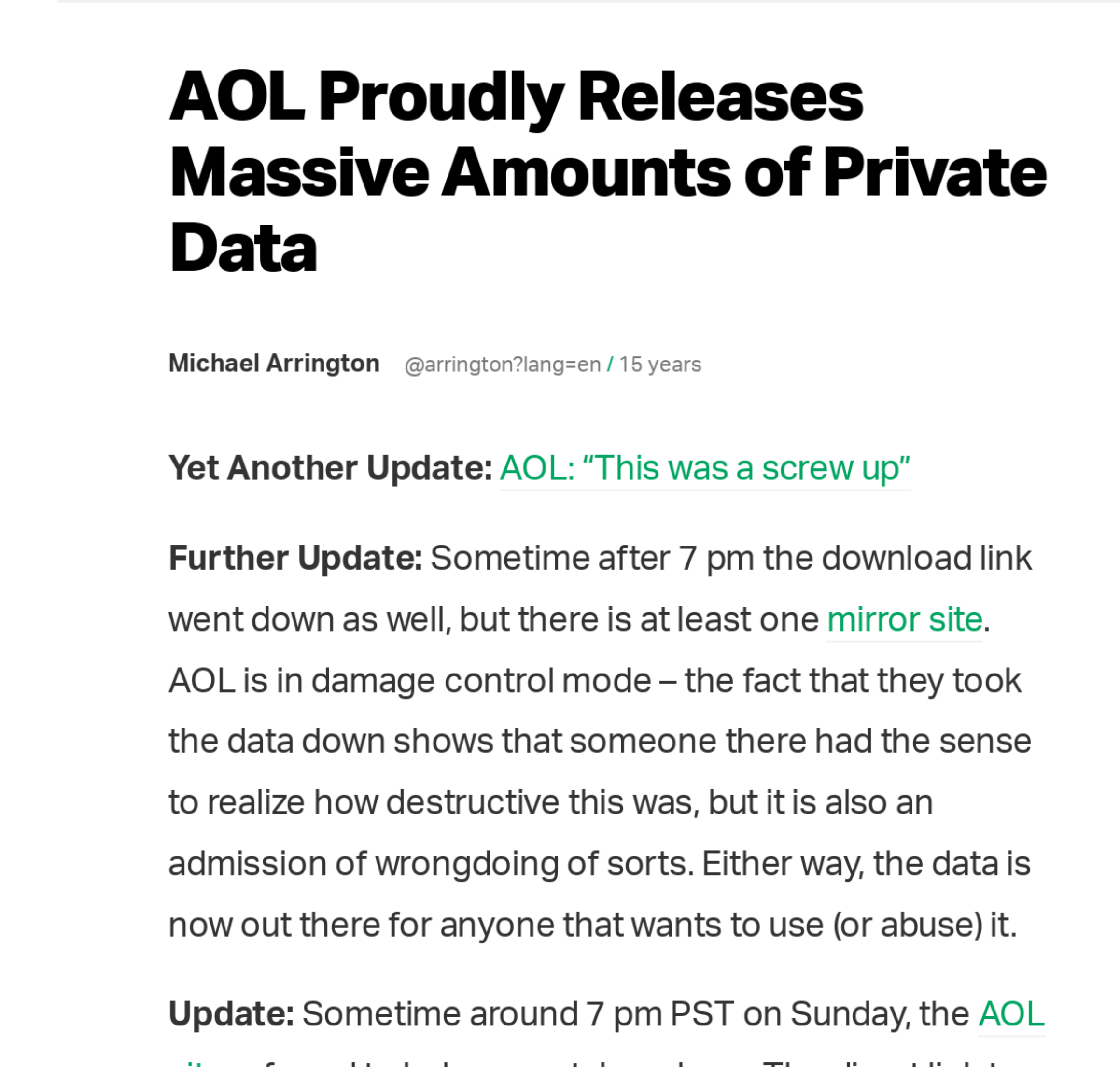
—Open Data Policy-Managing Information as an Asset
(http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf)

An example of the mosaic effect in humanitarian datasets: an edge indicates common column headers. Over 400 datasets were analyzed, over 90% (377) of datasets shared information with at least one other dataset.

Source: *Exploring the Mosaic Effect on HDX Datasets* (https://centre.humdata.org/exploring-the-mosaic-effect-on-hdx-datasets/)



In 2006, AOL released “anonymized” search data for research purposes…



AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 15 years

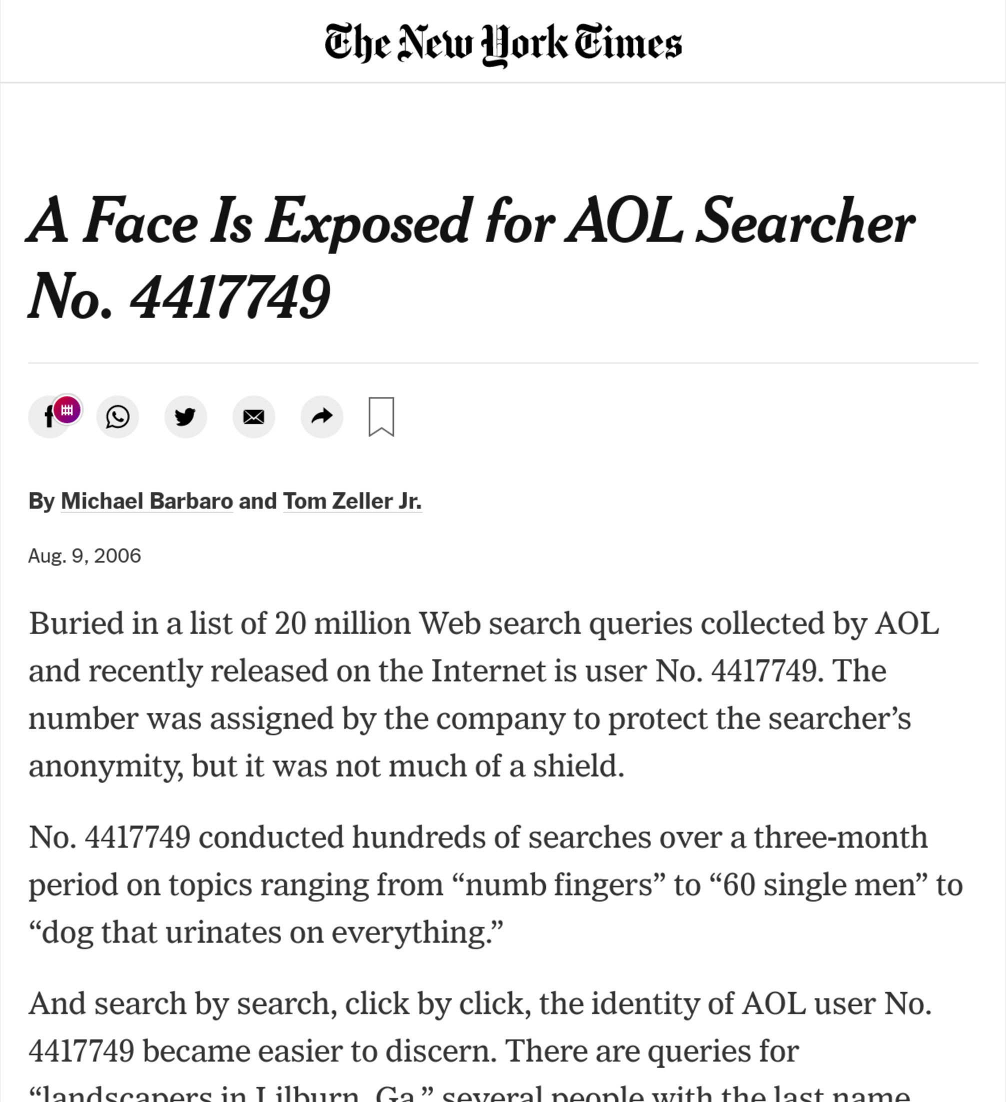
Yet Another Update: AOL: “This was a screw up”

Further Update: Sometime after 7 pm the download link went down as well, but there is at least one mirror site. AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.

Update: Sometime around 7 pm PST on Sunday, the AOL

Source: https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

The User ID in the data could be linked up with other information



The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name

Source: https://www.nytimes.com/2006/08/09/technology/09aol.html

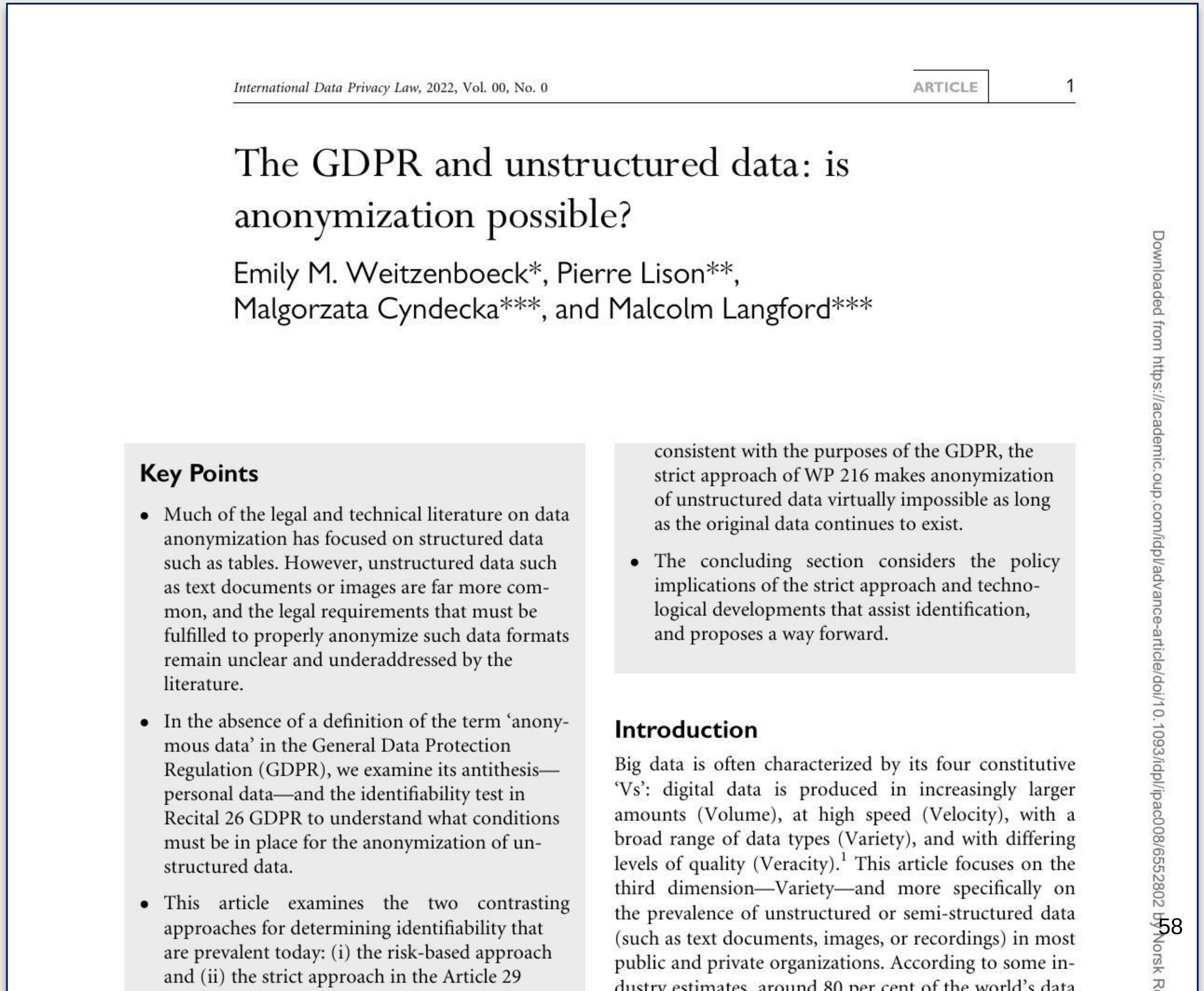
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem. Erik S. Lesser for The New York Times

Issue 4: Protecting raw data for good research practice

- To assist in peer-review and a possibility for helping in replication a study, raw data should be available on request.
- Keep the data, but pseudonymize the records using different numbers of real IDs. Keep raw data access restricted

Weitzenboeck and colleagues discuss the difficulties of anonymizing unstructured data

Source:
Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The GDPR and unstructured data: Is anonymization possible? *International Data Privacy Law*, ipac008. https://doi.org/10.1093/idpl/ipac008



International Data Privacy Law, 2022, Vol. 00, No. 0 ARTICLE 1

The GDPR and unstructured data: is anonymization possible?

Emily M. Weitzenboeck*, Pierre Lison**, Malgorzata Cyndecka***, and Malcolm Langford***

Key Points

- Much of the legal and technical literature on data anonymization has focused on structured data such as tables. However, unstructured data such as text documents or images are far more common, and the legal requirements that must be fulfilled to properly anonymize such data formats remain unclear and underaddressed by the literature.
- In the absence of a definition of the term 'anonymous data' in the General Data Protection Regulation (GDPR), we examine its antithesis—personal data—and the identifiability test in Recital 26 GDPR to understand what conditions must be in place for the anonymization of unstructured data.
- This article examines the two contrasting approaches for determining identifiability that are prevalent today: (i) the risk-based approach and (ii) the strict approach in the Article 29

consistent with the purposes of the GDPR, the strict approach of WP 216 makes anonymization of unstructured data virtually impossible as long as the original data continues to exist.
- The concluding section considers the policy implications of the strict approach and technological developments that assist identification, and proposes a way forward.

Introduction

Big data is often characterized by its four constitutive 'Vs': digital data is produced in increasingly larger amounts (Volume), at high speed (Velocity), with a broad range of data types (Variety), and with differing levels of quality (Veracity).[1] This article focuses on the third dimension—Variety—and more specifically on the prevalence of unstructured or semi-structured data (such as text documents, images, or recordings) in most public and private organizations. According to some industry estimates, around 80 per cent of the world's data

Weitzenboeck and colleagues present two ways of considering anonymization

1. Risk-based approach: how likely will a motivated intruder be able to reconstruct the data? The motivated intruder is not an elite hacker (1337 hax0r).
2. Working Party's Opinion on Anonymization Technique: data is traceable as long as the original source exists

Why Def. 2 is difficult: Example from Weitzenboeck et al: Original Data

1. The applicant [Mr Colin Joseph O'Brien] was born in 1955 and lives in Bridgend.
2. His wife died on 29 April 1999 leaving two children, born in 1989 and 1991.
3. In 1999 the applicant enquired about widows' benefits and he was informed that he was not entitled to such benefits.
4. In early 2000 the applicant applied for widows' benefits again and on 13 March 2000 the Benefits Agency rejected his claim.
5. He lodged an appeal against this decision on 16 March 2000 and this appeal was struck out on 23 May 2000 on the basis that it was misconceived.
6. On 16 May 2000 the applicant made an oral claim for Widow's Bereavement Allowance to the Inland Revenue. On 23 May 2000 he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.
7. The applicant received child benefit in the sum of GBP 100 per month.

Anonymous unstructured data is difficult: Data is masked …

1. The applicant [***] was born in *** and lives in ***
2. His wife died on *** leaving *** children, born in ***
3. In *** the applicant enquired about widows’ benefits and he was informed that he was not entitled to such benefits.
4. In *** the applicant applied for widows’ benefits again and on *** the *** rejected his claim.
5. He lodged an appeal against this decision on *** and this appeal was struck out on *** on the basis that it was misconceived.
6. On *** the applicant made an oral claim for Widow’s Bereavement Allowance to the Inland Revenue. On *** he was informed that his claim could not be accepted because there was no basis in domestic law allowing widowers to claim this benefit. The applicant was advised that an appeal against this decision would be bound to fail.
7. The applicant received child benefit in the sum of *** per month.

… but still traceable using the surrounding text!

Anonymous unstructured data is difficult: Fully anonymized so that it cannot be traced back to the original document. How useful is the data now?

1. The applicant [***] was born in *** and lives *** ***
2. *** *** *** *** *** *** two *** *** *** *** *** *** ***
3. In *** *** *** *** *** *** *** *** *** *** *** *** *** was *** *** *** *** *** ***
4. In *** the applicant *** *** *** *** *** *** *** *** the *** *** his *** ***
5. *** *** an *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** the *** that it was *** ***
6. *** *** *** *** *** *** *** *** for *** *** *** *** the *** *** *** *** *** *** *** *** *** *** *** could *** *** *** *** *** *** *** *** in *** law *** *** to *** this *** *** *** *** *** *** *** *** *** *** this *** *** *** *** to *** ***
7. The *** *** *** *** in the *** *** *** *** *** ***

In summary, researchers must protect the data they collect about people, but there is no universal solution

- The greater the vulnerability of the data subject, the larger the moral obligation of the researcher to protect the data subject from harm
- Harm depends on context and researchers need to have a good judgement about what can cause harm (assessing risk)

“When making ethical decisions, researchers must balance the privacy rights of the data subjects with the social benefits of the research and researchers’ rights to conduct research. In different contexts, the privacy rights of subjects may outweigh the benefits of research”

—Gisle Hannemyr