# ANONYMISERING AV PERSONOPPLYSNINGER

Veileder, 2015

Datatilsynet

# Innhold

# Innledning

Denne veilederen skal hjelpe virksomheter i arbeidet med å anonymisere innsamlede personopplysninger på en robust og sikker måte. Anonymisering handler om å fjerne muligheten for å identifisere enkeltpersoner i et datasett. Anonymisering er et viktig virkemiddel for å kunne hente ut verdifull innsikt ved dataanalyse, samtidig som risikoen reduseres for berørte personer. Når personopplysninger anonymiseres, regnes de ikke lenger som personopplysninger. Behandlingen av opplysningene faller da utenfor personopplysningsloven.

## Hvorfor denne veilederen

Å anonymisere data er utfordrende, og det er mer utfordrende i dag enn det var tidligere. Det enorme tilfanget av offentlig tilgjengelige data, kombinert med tilgang til stadig billigere og mer kraftfull analyseteknologi, har bidratt til å øke faren for reidentifisering.

Den økte risikoen for reidentifisering gjør det viktig å foreta grundige risikovurderinger før publisering av anonyme data, samt å benytte solide anonymiseringsteknikker. I denne veilederen vil vi gå gjennom sentrale juridiske bestemmelser, vise til risikofaktorer det er viktig å ta hensyn til, og diskutere styrker og svakheter ved ulike anonymiseringsteknikker.

## Hvem er veilederen for?

Veilederen er til for alle som ønsker å anonymisere personopplysninger. Rådene er relevante både for private og offentlige virksomheter, og uavhengig av formålet med anonymiseringen.  Det kan være mange årsaker til at en virksomhet ønsker å anonymisere innsamlede personopplysninger. Det kan for eksempel være fordi virksomheten

- er pålagt å publisere anonymiserte data
- må utgi informasjon til en tredjepart og ønsker å beskytte identiteten til de berørte
- ønsker å frigi data for å være åpen og transparent om egen virksomhet
- ønsker å bruke allerede innsamlede data til nye formål, for eksempel som grunnlag for profilbygging til bruk i målrettet markedsføring, eller for å se etter trender og mønstre
- ønsker å frigi data til bruk for statistiske formål eller til forskere for vitenskapelig bruk.

### Oppsummert

- Personvernlovgivningen gjelder ikke for anonyme data. Data er anonyme hvis det ikke lenger er mulig, med de hjelpemidlene som med rimelighet kan tenkes å ha blitt brukt, å identifisere enkeltpersoner i datasettet.
- Anonymisering av data gjør det mulig å utnytte verdien som ligger i dataanalyse på en personvernvennlig måte.
- Veilederen vil hjelpe virksomheter som ønsker å anonymisere personopplysninger, uavhengig av formål.
- Veilederen vil hjelpe virksomheter til å identifisere utfordringene og risikoene forbundet med å anonymisere personopplysninger slik at anonymiseringen blir så sikker som mulig.

# Anonymisering og personopplysninger

Når en person eller virksomhet med tilhold i Norge behandler digitale data, må vedkommende forholde seg til at behandlingen kan utløse plikter og rettigheter etter den norske personopplysningsloven. Den som behandler dataene må innrette seg etter de kravene loven stiller, ellers kan vedkommende risikere å måtte betale overtredelsesgebyr, bli erstatningsansvarlig, eller til og med havne i straffansvar.

Forutsetningen er imidlertid at behandlingen gjelder *personopplysninger*, ettersom personopplysningsloven bare gjelder opplysninger som angår enkeltpersoner. Personopplysningsbegrepets grenser blir derfor avgjørende for lovens anvendelse.

Litt forenklet kan man si at behandling av personopplysninger er omfattet av loven, mens behandling av anonyme eller anonymiserte data ikke er det. Det samme gjelder data som ikke knytter seg til personer i det hele tatt. Det er skillet mellom personopplysninger og anonyme data som er temaet i denne delen av veilederen.

## Personopplysninger og anonyme data

Hvor grensen går mellom personopplysninger og anonyme opplysninger, kan være vrient å avgjøre. Utgangspunktet for vurderingen er regelverkets definisjon av ordet personopplysning.

Personopplysninger er opplysninger og vurderinger som kan knyttes til en enkeltperson (personopplysningsloven § 2).

Denne definisjonen har tre hovedbestanddeler:

1. Det kan dreie seg om enhver form for informasjon
2. Tilknytningselementet *som kan knyttes til* er koblingsbroen mellom 1 og 3.
3. Identifiserbar eller identifisert enkeltperson

Vi tror en viss kjennskap til hva som er å anse som personopplysninger er nødvendig for å forstå anonymisering. For en grundigere analyse av personopplysningsbegrepet, se avsnitt 4.2 i rapporten Big Data – personvernprinsipper under press (2013, pdf), tilgjengelig fra datatilsynet.no.

En kort gjennomgang av de tre elementene:

### 1. Enhver form for informasjon

Alle typer informasjon er omfattet av definisjonen. For det første mener man *objektive* opplysninger. Dette kan være informasjon om en persons alder, bosted eller inntekt. For det andre kan det dreie seg om *subjektive* oppfatninger, for eksempel en persons vurderinger eller karakteristikker av en annen. Informasjonens sannhetsgehalt spiller ingen rolle, det er personopplysninger uansett om det dreier seg om en påstand, etterprøvbare fakta eller rent oppspinn.

Personopplysningsbegrepet er heller ikke begrenset til forhold som tradisjonelt forbindes med privatlivets fred. Også mer prosaiske forhold, som hvor man jobber eller hva man studerer, faller innenfor definisjonen av en personopplysning.

Spørsmålet om hvor beskyttelsesverdig informasjonen er, kommer først inn på et senere tidspunkt, gjerne når man skal vurdere om en behandling av de aktuelle opplysningene er i samsvar med loven eller ikke.

Det har heller ikke betydning hvilket format informasjonen kommer i. Personopplysninger kan uttrykkes i skrift, tall, tegninger, fotografier, lyd eller biometriske kjennetegn. Videre kan opplysningene finnes i e-poster, i offentlige saksdokumenter, på sosiale medier, i applikasjoner, sms-er, på nettet, og så videre.

### 2. Tilknytningselementet

Informasjonen må kunne knyttes til en fysisk person. Noen ganger er denne tilknytningen lett å konstatere, andre ganger er den det ikke. Opplysninger om tilstanden til en bil vil for eksempel trolig først og fremst forbindes med tingen i seg selv. Likevel kan den samme informasjonen også avsløre forhold om personer som har hatt befatning med tingen, for eksempel bilens eier. I visse tilfeller kan dessuten opplysninger om én person samtidig være opplysninger om en eller flere andre. Dette kan for eksempel være tilfellet i medisinsk eller genetisk sammenheng.

Dette er eksempler på at tilknytningen mellom informasjonen og personen kan være *indirekte*. En slik indirekte tilknytning er tilstrekkelig til at loven kan komme til anvendelse. Det fremgår direkte av ordlyden i personopplysningsloven § 2.

### 3. Identifiserbar enkeltperson

Informasjonen må kunne knyttes til en enkeltpersonen, og han eller hun må være *identifiserbar*. At en person er *identifisert* vil si at han eller hun er skilt ut fra en gruppe av personer. At personen er identifiser*bar*, innebærer at slik identifisering er *mulig*. At denne identifiseringen kan tenkes å finne sted på et eller annet tidspunkt i fremtiden, er tilstrekkelig.

Opplysninger kan ved første øyekast fremstå som anonyme, men likevel være personopplysninger i lovens forstand. Forklaringen er at det kan være mulig å identifisere en eller flere personer ad omveier. Et eksempel på dette er bilens registreringsnummer, eller en smarttelefons IMEI-nummer. Disse opplysningene kan i visse tilfeller kobles sammen med andre datasett eller

registre, og dermed avsløre identiteten til bilens og telefonens eier. Øvrige opplysninger som forekommer sammen med slike numre, for eksempel hvor bilen eller telefonen har vært, vil dermed også bli å regne som personopplysninger.

## Anonyme opplysninger

Over har vi forsøkt å forklare hva som ligger i personopplysningsbegrepet. Det er viktig å ha en viss forståelse av hva personopplysninger er, for å kunne avgjøre hva som skal til for at en personopplysning skal kunne regnes som anonymisert.

Personopplysningsloven stiller som nevnt ingen krav til behandling av anonyme eller anonymiserte opplysninger, og en korrekt identifisering av hvor grensen går kan dermed være av stor betydning.

Opplysninger av den typen som definert i punkt 1 over, kan sies å være anonyme når det ikke er mulig å finne et slikt tilknytningselement som i punkt 2, eller enkeltpersonen i punkt 3 ikke er identifiserbar.

Anonyme opplysninger kan defineres som opplysninger som det ikke er mulig å knytte til et identifiserbart individ, når man tar i betraktning alle de hjelpemidlene som med rimelighet kan tenkes brukt for å identifisere vedkommende, enten av den behandlingsansvarlige eller av en hvilken som helst annen person (Se avsnitt 26 i personverndirektivets fortale.)

### Definisjoner

**Anonymisering** er å gjøre personopplysninger anonyme.

**Pseudonymisering** vil si at enkelte direkte identifiserende parametere erstattes med pseudonymer, som fremdeles vil være unike indikatorer.

**Avidentifisering** vil si at alle personentydige kjennetegn er fjernet fra opplysningene, slik at de ikke lenger kan knyttes til en enkeltperson.

## Anonymisering

*Anonymisering* er å gjøre personopplysninger anonyme. Datasett som kan knyttes til en identifiserbar person bearbeides altså med tanke på å gjøre tilknytningen mellom informasjon og individ umulig. Flere teknikker kan brukes for å nå målet om å gjøre opplysningene anonyme. Svakheter og styrker ved de ulike teknikkene er beskrevet i vedlegget til denne veilederen (se side 16).

Når anonymiseringsprosessen er gjennomført, er det viktig å være klar over at det bare er snakk om reell anonymisering hvis prosessen er *irreversibel*. Det skal med andre ord ikke være mulig å gjenfinne koblingen mellom informasjon om enkeltindivid, når man tar i betraktning de hjelpemidlene som med rimelighet kan tenkes brukt for å identifisere vedkommende, slik som nevnt i avsnittet over.

Vurderingen av om opplysningene gjør det mulig å identifisere en person, og om opplysningene kan betraktes som anonyme eller ei, avhenger av de faktiske omstendighetene. Man må ta utgangspunkt i en vurdering av hvor sannsynlig muligheten for reidentifisering er. De enkelte tilfellene må vurderes og analyseres på bakgrunn av hvilke hjelpemidler som finnes i dag, men også med tanke på morgendagens hjelpemidler – innenfor rimelighetens grenser, naturligvis. Målestokken er hvorvidt disse hjelpemidlene med rimelighet kan tenkes brukt til å finne identiteten til de involverte.

Av og til forveksles anonymisering med to lignende fenomener, nemlig *pseudonymisering* og *avidentifisering*. Slik forveksling kan være uheldig, i verste fall kan det resultere i at man gjør seg skyldig i lovbrudd med de følgene det kan få.

## Pseudonymisering

*Pseudonymisering* vil si at enkelte direkte identifiserende parametere erstattes med pseudonymer, som fremdeles vil være unike identifikatorer. Det eksisterer derfor en sannsynlighet for at enkeltindividet vil kunne bli indirekte identifisert, og ofte er det et poeng at man kan følge den samme (pseudonymiserte) personen over en viss tid, for eksempel i forbindelse med forskning. Dermed ser vi at vi befinner oss innenfor lovens personopplysningsbegrep, og konsekvensen av det blir at lovens regler må respekteres.

Det er med andre ord meget viktig å være oppmerksom på denne forskjellen, ettersom pseudonymiserte opplysninger altså er underlagt de kravene som personopplysningsloven stiller, mens det motsatte er tilfellet når man har å gjøre med anonyme data.

Det er ikke dermed sagt at pseudonymisering er uten fordeler. Pseudonymisering kan gjøre det vanskeligere å knytte et bestemt sett av data til den registrertes identitet, og følgelig kan pseudonymisering ses på som et nyttig tiltak som fremmer personvernet. Pseudonymiseringen vil kunne beskytte det individet som opplysningen knytter seg til, og det kan være lettere å få forankret behandlingen av de pseudonymiserte opplysninger i ett eller flere av lovens rettslige grunnlag.

Begrepene vi bruker i denne veilederen baserer seg på felles europeiske vurderinger (se for eksempel Artikkel 29-gruppens uttalelse om anonymiseringsteknikker (pdf, engelsk)). De kan avvike fra slik begrepene er innarbeidet i Norge, særlig i helsesektoren. I helseregisterloven som trådte i kraft 1. januar 2015 ble de tidligere brukte definisjonene på pseudonymiserte og avidentifiserte helseopplysninger erstattet med begrepet «indirekte identifiserbare helseopplysninger», som er bredere. (Mer om begrepet pseudonymisering slik det var å forstå før vi fikk ny lov, finnes i rundskriv I-8/2005 (regjeringen.no, pdf). Dette gjelder også begrepet kryptering i en rettslig forstand, som avviker fra hvordan begrepet er brukt som en teknisk metode i denne veilederen).

## Fordelene med å anonymisere data

Har du et tilstrekkelig robust og sikkert anonymisert datasett, kan du nyttiggjøre deg informasjonen uten å risikere å komme på kant med personopplysningsloven. Du behøver ikke å ta hensyn til spørsmål om *behandlingsansvar,* og videre bruk og analyse av denne typen data er ikke underlagt noen konsesjonsplikt eller meldeplikt.

Du behøver heller ikke forsikre deg om at det foreligger noe rettslig grunnlag for behandlingen av dem, du trenger ikke forholde deg til krav om relevans eller formålsbegrensning, og den som sitter på dataene er ikke underlagt forpliktelse til å slette opplysningene, og så videre.

Anonymisering kan være løsningen i de tilfellene der det kan herske tvil om loven tillater en bestemt behandling av personopplysninger. I tillegg kommer tilfeller hvor behandlingen av personopplysninger er utelukket, fordi lovens krav setter en stopper for den. Anonymiserer man dataene, er man utenfor loven.

## Anonymisering og behandlingsbegrepet

Det er for øvrig en forutsetning at opplysningene som skal anonymiseres er samlet inn og behandlet i samsvar med personopplysningslovens krav. I teorien må også selve anonymiseringen regnes som en behandling av personopplysninger. Konsekvensen er at den som anonymiserer opplysningene, må respektere kravene i personopplysningsloven § 11 underveis i anonymiseringsprosessen (se pkt 2.2.1 i Artikkel 29-gruppens uttalelse om anonymiseringsteknikker (pdf, engelsk) for mer.) Formålsbegrensningen i § 11 første ledd bokstav c må for eksempel respekteres.

Anonymiseringen vil trolig ofte kunne hjemles i den såkalte interesseavveiningsbestemmelsen i personopplysningsloven § 8 bokstav f. Bestemmelsen sier at personopplysninger bare kan behandles dersom behandlingen er nødvendig for at den behandlingsansvarlige, eller tredjepersoner som opplysningene utleveres til, kan ivareta en berettiget interesse, *og* hensynet til den registrertes personvern ikke overstiger denne interessen.

Det må altså foreligge en berettiget interesse, og lovens nødvendighetskrav må være oppfylt. Et sentralt poeng er imidlertid at den registrertes personvern i svært liten grad kan sies å bli krenket gjennom en anonymisering av opplysninger som kan knyttes til ham eller henne. Dette vil naturligvis spille en viktig rolle i avveiningen.

Regner man anonymisering som en behandling i lovens forstand, ser man at anonymisering ikke kan «reparere» manglende legalitet eller legitimitet ved den opprinnelige datainnsamlingen. Man kommer altså ikke unna med å samle inn personopplysninger i strid med loven i første omgang, for så å anonymisere i neste.

# Utfordringer knyttet til anonymisering

## Risiko for reidentifisering

Den enorme tilgangen til offentlig tilgjengelige data, kombinert med stadig mer kraftfull analyseteknologi, har bidratt til å øke risikoen for reidentifisering. Reidentifisering innebærer at man klarer å identifisere enkeltpersoner fra i utgangspunktet antatt anonyme datasett. Studier har vist at man ved å sammenstille data fra flere kilder kan reidentifisere personer ved kun å kjenne til to attributter i et anonymisert datasett, som for eksempel postnummer og fødselsdato.

Reidentifisering kan skje ved at noen tar personlige data de allerede har om andre og søker etter treff i et anonymisert datasett, eller ved at man tar et treff fra et anonymt datasett og søker etter treff på offentlig tilgjengelig informasjon. Eksempler på slik offentlig tilgjengelig informasjon kan være data fra offentlige registre (for eksempel skattelisten, Brønnøysundregistrene, kjøretøyregisteret, Offentlig elektronisk postjournal), sosiale medier, lokale og nasjonale pressearkiv og nettsteder for slektsforskning.

Det finnes flere kjente tilfeller av reidentifisering, de fleste gjennomført av forskere på reelle datasett. Felles for de fleste av disse er at dataene i utgangspunktet var dårlig anonymisert. Virksomhetene har for eksempel beholdt for mange identifiserende elementer i datasettet, eller har ikke i forbindelse med publiseringen av dataene kartlagt hvilke andre tilgjengelige datasett som finnes «der ute» som kan bruke til å utlede informasjon fra deres datasett.

Enkelte datatyper er mer utfordrende å anonymisere enn andre, for eksempel lokasjonsdata og genetiske

### Eksempel: Netflix

Et kjent eksempel på et tilfelle av publisering av dårlig anonymiserte data er det om da Netflix annonserte en konkurranse for utviklere. Premien var på én million amerikanske dollar. Målet var at noen skulle utvikle en løsning som ga en forbedring på 10 prosent på deres anbefalingsmodul.

I den forbindelse slapp Netflix et «øvingsdatasett» til de konkurrerende utviklerne som de kunne bruke for å trene sine systemer. Med datasettet fulgte en ansvarsfraskrivelse (disclaimer): «For å beskytte kundenes personvern, har all personlig informasjon som identifiserer den enkelte kunde blitt fjernet, og alle kundens id-er har blitt erstattet med tilfeldig tildelte id-er.»

Det finnes flere filmvurderingsportaler på Internett, blant annet IMDb. På IMDb kan enkeltpersoner registrere seg og rangere filmer, og stå frem med fullt navn.

Forskerne Narayanan og Shmatikov koblet Netflix' avidentifiserte treningsdatabase med IMDbs database (basert på datoen for vurdering av en bruker) og klarte på den måten delvis å reidentifisere brukerne i Netflix' øvingsdatabase.

### Anonymisering oppsummert

- Personopplysninger er opplysninger og vurderinger som kan knyttes til en enkeltperson.
- Det kan være vanskelig å avgjøre hvor grensen går mellom personopplysninger og anonyme opplysninger. Det er derfor viktig å ha en viss forståelse av hva personopplysninger er.
- Anonyme opplysninger faller utenfor personopplysningslovens virkeområde. For at personopplysningsloven ikke skal gjelde, er det avgjørende at anonymiseringen er reell. Det vil si at det ikke skal være mulig å gjenfinne koblingen mellom informasjon og enkeltindivid, tatt i betraktning hvilke hjelpemidler som med rimelighet kan tenkes å ha blitt brukt.
- Fordelen med anonymisering er at den videre behandlingen av dataene kan skje uten noen form for behandlingsansvar.
- Anonymisering vil ikke alltid være nødvendig. I mange tilfeller vil den aktuelle databehandlingen være forankret i et eller flere av lovens rettslige grunnlag.

opplysninger. Menneskets bevegelsesmønster er så unikt at den semantiske delen av lokaliseringsdataene – stedene der den registrerte har vært på et bestemt tidspunkt – kan avsløre mange detaljer om en registrert, selv uten andre kjente verdier for attributter. Dette er blitt påvist i mange representative akademiske undersøkelser (de Montjoye et al: «Unique in the Crowd: The privacy bounds of human mobility», Nature, 3, 1376 (2013)).

### Eksempel: AOL

Et typisk eksempel på misforståelsene omkring pseudonymisering er AOL-saken.

I 2006 ble en database med 20 millioner søkeord for over 650 000 brukere offentliggjort. Det eneste AOL hadde gjort for å ta hensyn til personvernet til sine brukere, var utskiftning av AOL-brukernavnet med et numerisk løpenummer. Dette resulterte i at noen av personene ble identifisert og lokalisert.

Pseudonymiserte søkestrenger for søkemotorer har en meget høy identifiseringskraft, særlig hvis de sammenkobles med andre attributter som ip-adresser eller andre klientkonfigurasjons-parametre.

Genetiske dataprofiler er et annet eksempel på personopplysninger som risikerer å bli reidentifisert hvis den eneste anonymiseringsteknikken som brukes er å fjerne donorens identitet. Dette fordi genene i seg selv er helt unike. Det er blitt påvist i studier at kombinasjonen av offentlig tilgjengelige genetiske ressurser (for eksempel slektsregister, nekrologer, søkeresultater fra søkemotorer) og metadata om DNA-donorer (donasjonstidspunkt, alder, bosted) kan avsløre visse personers identitet, selv om de har avgitt DNA «anonymt».

## Pseudonyme data betraktes som anonyme data

Pseudonymiserte data er ikke ensbetydende med anonymiserte data, som poengtert i kapittel 2. Pseudonymisering åpner for identifisering av enkeltpersoner. Behandlingsansvarlige som velger å pseudonymisere opplysningene heller enn å anonymisere dem, må være klar over at opplysningene da fortsatt vil være definert som personopplysninger og derfor må behandles etter kravene i personopplysningsloven.

Det finnes flere eksempler på at behandlingsansvarlige tror de har anonymisert dataene, mens de i virkeligheten kun har pseudonymisert opplysningene ved for eksempel å bytte ut navnene med et løpenummer.

## Kryptering er ikke anonymisering

En annen utbredt misforståelse er å sette likhetstegn mellom krypterte og enveiskodede (hashede) data, og anonymiserte data. Denne misforståelsen beror på to antagelser, nemlig at a) når et element i en database (for eksempel navn, adresse, fødselsdato) er blitt kryptert eller erstattet av en randomisert kode ved hjelp av krypteringsteknologi, for eksempel en hashfunksjon med nøkkel, så er denne posten anonymisert, og b) at anonymiseringen blir mer effektiv hvis nøkkelens lengde er riktig og man bruker en avansert krypteringsalgoritme.

Det er viktig å være klar over at formålene med kryptering og anonymisering er forskjellig: Kryptering er en sikkerhetsmetode som har til formål å sikre fortroligheten i en kommunikasjonskanal mellom to identifiserte parter (mennesker, enheter eller programvare) for å unngå avlytting eller utilsiktet offentliggjøring. Formålet med anonymisering er på den annen side å unngå at personer blir identifisert ved å forhindre skjult sammenkobling av attributter knyttet til den registrerte personen.

Verken kryptering eller nøkkelkoding som sådan bidrar til å gjøre den registrerte uidentifiserbar, ettersom de originale dataene fremdeles er tilgjengelige eller kan utledes i hvert fall hos den behandlingsansvarlige. Å kun foreta en semantisk oversetting av personopplysningene, som tilfellet er med nøkkelkoding, fjerner ikke muligheten til å gjenskape dataene til deres opprinnelige struktur.

Avansert kryptering kan bidra til å beskytte dataene bedre gjennom å gjøre dem uforståelige for aktører som ikke har tilgang til krypteringsnøkkelen, men det innebærer ikke nødvendigvis at dataene er anonyme. Så lenge nøkkelen til de originale dataene er tilgjengelig (selv om den oppbevares av en betrodd tredjepart), fjernes ikke muligheten til å identifisere den registrerte. Krypterte og enveiskodede data er derfor å betrakte som personopplysninger og må behandles deretter.

## Kryptering vs. anonymisering

Kryptering blir brukt for å sikre fortrolighet i en kommunikasjonskanal, og har et annet formål enn å anonymisere. Formålet med anonymisering er å unngå at personer blir identifisert.

(Dette kapittelet er basert på og delvis en oversettelse av Artikkel 29-gruppens uttalelse om anonymiseringsteknikker (pdf))

### Utfordringer oppsummert

- Fare for reidentifisering: ved sammenstilling av data fra flere kilder kan det oppstå risiko for at enkeltindivider kan identifiseres fra i utgangspunktet anonyme datasett.
- To kjente datapunkter, for eksempel postnummer og fødselsdato, kan være nok til å reidentifisere enkeltpersoner ut fra datasettet.
- Pseudonyme data må ikke forveksles med anonyme data. Pseudonyme data er personopplysninger, og behandling av slike data faller inn under personopplysningsloven.
- Kryptering er ikke det samme som anonymisering. Formålet med kryptering er å beskytte data, ikke å gjøre dem uidentifiserbare.

# Anbefalinger for sikker anonymisering

## Skal opplysningene anonymiseres, pseudonymiseres eller være identifiserbare?

Den behandlingsansvarlige må på et tidlig tidspunkt beslutte om personopplysningene som skal behandles skal anonymiseres, avidentifiseres eller om de skal være identifiserbare. Dette valget påvirker hvordan virksomheten må forholde seg til personopplysningsloven i den videre behandlingen av opplysningene. Velges anonymisering, vil som nevnt den videre prosessen falle utenfor personopplysningslovens virkeområde.

## Beskriv formålet med anonymiseringen

Det er viktig å ha klart for seg hva formålet med anonymiseringen er. Bruksområdet til datasettet spiller en avgjørende rolle med hensyn til å fastslå risikoen for reidentifisering. Hvis opplysningene skal publiseres offentlig på Internett, innebærer dette en større risiko enn begrenset frigivelse av data, for eksempel forskningsformål til eller til bruk i egen virksomhet. Selv om begrenset frigivelse er enklere å kontrollere og vurdere, er dette allikevel heller ikke uten risiko.

Flere forhold må vurderes når opplysninger skal anonymiseres, for eksempel:

- Hvilken type data er det som skal anonymiseres?
- Hvilke kontrollmekanismer er knyttet til datasettet? Hvilke sikkerhetsforanstaltninger skal begrense tilgangen til dataene?
- Størrelsen på datasettet (Hvilke kvantitative egenskaper har det?)
- Hvilke andre offentlige tilgjengelige informasjonsressurser finnes som potensielt kan brukes til å utlede informasjon om enkeltpersoner i det datasettet som skal anonymiseres?
- Skal datasettet frigis til tredjeparter? (Skal det gjøres i begrenset form, eller gjøres det offentlig tilgjengelig, for eksempel på Internett)?
- Finnes det utenforstående aktører som kan tenkes å ville foreta et målrettet angrep på dataene for å forsøke å identifisere enkeltpersoner? (I denne typen risikovurdering har særlig opplysningenes følsomhet og type betydning.)

## Foreta en risikovurdering før og etter publisering

Det er nesten umulig å vurdere risikoen for reidentifisering med absolutt sikkerhet. Å få absolutt oversikt over all annen informasjon som «*er der ute*», hvem den er tilgjengelig for og hvorvidt den kan komme til å bli brukt i et reidentifiseringsforsøk, er svært utfordrende. Slik «annen informasjon» kan være informasjon som er tilgjengelig for en annen organisasjon eller bestemte personer, eller informasjon som er allment tilgjengelig på Internett.

Ettersom man aldri helt kan vite hvilke data som er tilgjengelig i øyeblikket, eller som vil bli gjort tilgjengelig en gang i fremtiden, er det nødvendig å foreta en så grundig risikoanalyse som mulig tidlig i anonymiseringsprosessen.

Det er videre viktig at den behandlingsansvarlige ikke glemmer å tenke risiko rundt et datasett også *etter* at det er publisert. På grunn av risikoen for reidentifisering bør den behandlingsansvarlige *regelmessig* undersøke om det har dukket opp *nye* risikofaktorer og *re-evaluere* risikofaktorene som allerede er identifisert. Det er også viktig å vurdere om *kontrollen* med de identifiserte risikofaktorene er tilstrekkelig og om nødvendig justere den.

Hvis noen skulle lykkes i å reidentifisere opplysninger, og dette resulterer i at personopplysninger blir behandlet, må virksomheten som er ansvarlig for opplysningene påta seg behandlingsansvar for dem etter personopplysningslovens bestemmelser.

## Bruk en reidentifiseringstest

En test som kan være nyttig å utføre for å teste risikoen knyttet til reidentifisering, er den såkalte «motivert inntrenger»-testen (the «motivated intruder»-test). Denne testen innebærer at man tester om dataene lar seg reidentifisere *hvis* en inntrenger skulle prøve på dette.

Den motiverte inntrengeren skal vurderes som en person/organisasjon som uten forutgående kunnskaper prøver å identifisere individer i et anonymisert datasett. Selv om vedkommende ikke har forutgående kunnskaper, skal han betraktes som tilstrekkelig kompetent. Det vil si at vedkommende har tilgang til ressurser som Internett, offentlige registre og biblioteker. Det skal også forutsettes at han vil være i stand til å bruke ulike etterforskningsteknikker for å få i tale folk med kunnskap om identiteten til enkeltpersoner i datasettet.
Den motiverte inntrengeren skal imidlertid ikke vurderes til å ha spesialistkunnskaper, som for eksempel ekspertise i datahacking, eller tilgang til spesialutstyr for å bryte seg inn for å få tilgang til data som er sikkert oppbevart.

Enkelte typer data er opplagt mer attraktive for en motivert inntrenger enn andre. Det kan være mange årsaker til at noen vil forsøke å reidentifisere enkeltpersoner i et anonymisert datasett, for eksempel:

- for å avsløre nyhetsverdig informasjon om offentlige personer
- av politiske eller aktivistiske formål, for eksempel som del av en kampanje mot en bestemt organisasjon eller en person
- av nysgjerrighet, et ønske om å finne ut hvem som er Er det mulig å *utlede* informasjon knyttet til en enkeltperson fra datasettet?
- involvert i en lokal byggesak for eksempel.

### Motivert inntrenger-test – en sjekkliste

- Hva er risikoen for et såkalt «jigsaw attack», det vil si at man setter sammen ulike biter av informasjon som til sammen danner et mer komplett bilde av en person? Er dataene av en slik art at de kan lenkes sammen – er for eksempel samme kode brukt for å referere til samme individ på tvers av ulike datasett?
- Hvilken annen type «lenkbar» informasjon finnes lett eller offentlig tilgjengelig?
- Hvilke tekniske tiltak kan benyttes for å lykkes i å reidentifisere dataene?
- Hvor mye vekt skal man tillegge enkeltpersoners mulige kunnskap om de registrerte i det anonymiserte datasettet?
- Hvis en inntrenger-test utføres, hvilke svakheter avslørte den?

**Informasjonskilder kan være** biblioteker, offentlige registre (skattelister, Brønnøysundregistret, Kjøretøyregistret), elektronisk postjournal (oversikt over inngående og utgående korrespondanse i offentlig forvaltning), kirkebøker, slektsforsknings-nettsider, sosiale medier, søkemotorer, lokale og nasjonale pressearkiv, anonymiserte data publisert av andre organisasjoner.

Dette betyr imidlertid ikke at data som ved første øyekast fremstår som «ordinære», «ufarlige» eller uten verdi kan bli publisert uten en grundig vurdering av trusselen knyttet til reidentifikasjon.

Motivert inntrenger-testen er nyttig fordi den setter terskelen for risiko for reidentifisering høyere enn til hvorvidt en vanlig ukyndig borger skal kunne klare å reidentifisere dataene, men lavere enn til hvorvidt en ekspert med tilgang til spesialkompetanse, analytisk kraft og forkunnskaper skal kunne utføre det. Det er god praksis å foreta en motivert inntrenger-test som del av en risikovurdering. I praksis vil gjennomføringen av en motivert inntrenger-test for eksempel kunne omfatte:

- å gjennomføre et nettsøk for å finne ut om en kombinasjon av fødselsdato og postnummer kan brukes til å avsløre et enkeltindivids identitet
- å søke i arkivene til en nasjonal eller lokal avis for å se om det er mulig å knytte offerets navn med data over hvor ugjerninger er begått
- å bruke sosiale nettsamfunn for å se om det er mulig å knytte anonymiserte data til en brukers profil
- å bruke data fra ulike offentlige registre for å prøve å koble anonymiserte data til noens identitet

Ettersom tilgangen til tilgjengelig informasjon og datakraft hele tiden øker, er det viktig å regelmessig foreta en motivert inntrenger-test for å revurdere risikoen for reidentifisering.

## Anonymiseringsteknikk må bestemmes i hvert tilfelle

Det finnes forskjellige anonymiseringsteknikker, som er mer eller mindre robuste. Den behandlingsansvarlige må ta hensyn til hvilken garanti mot reidentifisering som kan oppnås ved å bruke en bestemt teknikk, vurdert ut fra tre sentrale risikoer forbundet med anonymisering:

- Er det fortsatt mulig å *skille ut* en enkeltperson i datasettet?
- Er det mulig å *koble sammen* ulike datasett knyttet til en og samme person?
- Er det mulig å *utlede* informasjon knyttet til en enkeltperson fra datasettet?

Ingen anonymiseringsteknikker oppfyller kravene til effektiv anonymisering fullt ut. Hvordan et datasett best skal anonymiseres må derfor avgjøres i hvert enkelt tilfelle, der man tar hensyn til formålet med anonymiseringen og konteksten. Det er nødvendig å være bevisst og omhyggelig i arbeidet med å bestemme hvilken teknikk som passer best i en gitt situasjon, samt hvordan flere teknikker eventuelt kan brukes *i kombinasjon* for å gjøre resultatet enda mer robust.

| | **Er det fortsatt mulig å *skille ut* en enkeltperson i datasettet?** | **Er det fortsatt mulig å *koble sammen* ulike datasett knyttet til en og samme person?** | **Er det fortsatt mulig å *utlede* informasjon knyttet til en enkeltperson?** |
|---|---|---|---|
| **Tilføring av støy** | Ja | Trolig ikke | Trolig ikke |
| **Substitusjon** | Ja | Ja | Trolig ikke |
| **Aggregering eller K-anonymitet** | Nei | Ja | Ja |
| **L-diversitet** | Nei | Ja | Trolig ikke |
| **Differential privacy** | Kanskje/trolig ikke | Trolig ikke | Trolig ikke |
| **Hashing (tokenisering)** | Ja | Ja | Trolig ikke |
| **Pseudonymisering** | Ja | Ja | Ja |

Kjenner man til svakhetene og styrkene til de ulike teknikkene, er det lettere å avgjøre hvordan man kan anonymisere opplysningene på en mest mulig sikker måte. I vedlegg 2 er svakheter og styrker til de ulike teknikkene gjennomgått i detalj.

Tabellen ovenfor gir en oversikt over styrker og svakheter til de ulike metodene i forhold til de tre grunnleggende kriteriene nevnt over. En løsning som oppfyller alle de tre kriteriene vil være motstandsdyktig mot identifisering. Hvis et løsningsforslag ikke oppfyller et av kriteriene, bør det utføres en grundig evaluering av identifiseringsrisikoene til/i datasettet.

(Dette kapittelet er basert på og delvis en oversettelse av innhold i to publikasjoner: Artikkel 29-gruppens uttalelse om anonymiseringsteknikker og en veileder for anonymisering av personopplysninger utgitt av Information Commissioner's Office, Datatilsynets britiske søsterorganisasjon: Anonymisation: managing data protection risk code of practice, ICO (2012). Vi gjenbruker teksten med tillatelse fra ICO.)

## Veilederen oppsummert

- Velg anonymiseringsteknikk(er) ut i fra kontekst og formål.
- Den optimale anonymiseringsløsningen må velges fra gang til gang.
- Alle teknikker har sine svakheter og styrker. Bruk derfor gjerne en kombinasjon av teknikker for å sikre best mulig anonymisering.
- Risikovurderinger må foretas før publisering av dataene, men også etter at dataene er publisert. Ny risiko kan komme til.
- Bruk tester for å vurdere sannsynligheten for reidentifisering.
- Den behandlingsansvarlige bør opplyse om hvilke anonymiseringsteknikker, eller kombinasjoner av teknikker, som er brukt i forbindelse med publisering av anonymiserte datasett.
- Lett identifiserbare (for eksempel sjeldne) attributter bør fjernes fra datasettet.

# Ordliste

| | |
|---|---|
| **Aggregerte data** | Statistiske data om flere individer som har blitt kombinert for å vise generelle trender eller verdier uten å identifisere enkeltindivider i datasettet. Aggregerte data er ikke nødvendigvis anonyme data. |
| **Angriper** | En tredjeperson (dvs. verken den behandlingsansvarlige eller en databehandler), som tilegner seg tilgang til de originale oppføringene, enten tilfeldig eller forsettlig. |
| **Anonyme data** | Data som det ikke er mulig, ved hjelp av alle rimelige tekniske hjelpemidler, å knytte tilbake til en enkeltperson. |
| **Anonymisering** | Prosessen med å gjøre om data slik at enkeltindivider ikke lengre blir identifisert og der faren for reidentifisering ikke er tilstede, ved hjelp av alle rimelige tekniske hjelpemidler. |
| **Attributt** | I en relasjonsdatabase er et attributt en egenskap til tabellene (entitetene). For eksempel i et kunderegister har man en tabell (entitet) som heter Kunde. Da kan man ha attributter som heter kundenummer, fornavn, etternavn, gateadresse, postnummer mv. I en tabell vil attributtene være navn på de ulike kolonnene. Verdier tilordnet attributtene kan da være 123456, Per, Hansen, Storgata 1, 0123. |
| **Avidentifisert** | Data hvor direkte identifiserende verdier for attributter har blitt fjernet og erstattet med en unik identifikator for å skjule identiteten til de registrerte. |
| **Begrenset tilgang** | Frigjøring av data for en begrenset og nærmere definert krets, for eksempel for forskere eller en institusjon. |
| **Behandlingsansvarlig** | Den som bestemmer formålet med behandlingen av personopplysninger og hvilke hjelpemidler som skal brukes. Vedkommende har ansvaret for at opplysninger behandles i henhold til kravene i personopplysningsloven. |

| | |
|---|---|
| **Databehandler** | Den som behandler personopplysninger på vegne av den behandlingsansvarlige. Vedkommende skal bare behandle personopplysninger i henhold til avtale med den behandlingsansvarlige. |
| **Datasett** | Et datasett består av forskjellige oppføringer relatert til personer (registrerte). Hver enkelt oppføring er relatert til én registrert og består av et sett verdier (f.eks. 2013) for hver attributt (f.eks. år). Et datasett er en samling oppføringer, som kan utformes som en tabell (eller en rekke tabeller) eller som en annotert/vektet graf. |
| **Kvasiidentifikatorer** | Kombinasjoner av oppføringer relatert til en registrert eller en gruppe registrerte. I noen tilfeller kan et datasett ha flere oppføringer om samme person. |
| **Oppføring** | Alt innholdet i en rad i en tabell (f.eks. Per Hansen, Rødveien 9, osv). (i motsetning til attributt som er navn (tittel) på en kolonne (f.eks. Fornavn)). |
| **Personopplysning** | Er en opplysning eller vurdering som kan knyttes til deg som enkeltperson |
| **Pseudonymisering** | En prosess der direkte identifiserende parametere erstattes med unike indikatorer for ikke å avsløre den virkelige identiteten til de registrerte. |
| **Publisering** | Allmenn tilgjengeliggjøring av data, for eksempel gjennom å publisere dem på Internett. |
| **Registrerte** | Enkeltindivider som er oppført i et register. |
| **Sensitive personopplysninger** | Opplysninger om rasemessig eller etnisk bakgrunn, eller politisk, filosofisk eller religiøs oppfatning, at en person har vært mistenkt, siktet, tiltalt eller dømt for en straffbar handling, helseforhold, seksuelle forhold eller medlemskap i fagforeninger. |

# Vedlegg: Svakheter og styrker ved ulike anonymiseringsteknikker

## Tre sentrale risikoer ved anonymisering

Det finnes forskjellige anonymiseringsteknikker, som er mer eller mindre robuste. Dette avsnittet handler om de viktigste punktene som den behandlingsansvarlige må vurdere i bruken av disse. Den behandlingsansvarlige må ta hensyn til hvilken garanti som kan oppnås ved å bruke en bestemt teknikk vurdert ut ifra tre sentrale risikoer forbundet med anonymisering:

- *Utskillelse* (singling out): muligheten for å isolere eller skille ut noen eller alle oppføringer, som identifiserer en enkeltperson i et datasett.
- *Sammenkobling* (linkability): muligheten for å koble sammen minst to oppføringer tilhørende den samme registrerte eller en gruppe registrerte (enten i samme database eller i to forskjellige databaser). Hvis en angriper (f.eks. ved hjelp av korrelasjonsanalyse) kan identifisere to oppføringer tilknyttet samme *gruppe* personer, men ikke enkeltpersoner i selve gruppen, beskytter teknikken mot utskillelse, men ikke mot sammenkobling.
- *Uttrekking* (inference): muligheten for at man med stor sannsynlighet kan trekke ut verdien av et attributt fra verdiene av et sett andre attributter, hvilket gir en mulighet for å utlede identifiserende opplysninger om den registrerte.

En løsning som beskytter mot alle tre risikoene vil være robust overfor reidentifisering. Det er imidlertid viktig å understreke at det løpende blir forsket på avidentifiserings- og anonymiseringsteknikker, og at denne forskningen til stadighet viser at ingen teknikker er uten svakheter

Det finnes i hovedsak to forskjellige fremgangsmåter for anonymisering: Den første er basert på *randomisering, mens den andre er basert på generalisering*.

## Randomisering

Randomisering er en gruppe teknikker som endrer dataenes nøyaktighet, slik at forbindelsen mellom data og person fjernes. Hvis dataene gjøres tilstrekkelig upresise, kan de ikke lenger vise til en bestemt person. Randomisering vil i seg selv ikke redusere de enkelte oppføringenes individualitet, da de stadig vil stamme fra en registrert, men kan beskytte mot angrep/risikoer basert på uttrekking. Det kan være nødvendig å også benytte generaliseringsteknikker i tillegg til randomisering, for å øke beskyttelsen av de registrerte. Følgende typer teknikker hører inn under randomisering:

**Å tilføre støy («noise addition»)**

Denne teknikken er særlig anvendelig hvis man behandler spesielt sensitive data, og der en eventuell reidentifisering kan få store konsekvenser for de registrerte. Den består i å endre attributter i datasettet, så de ikke er helt nøyaktige, samtidig med at den overordnede fordelingen opprettholdes. Når observatører behandler et datasett, vil de anta at verdiene er nøyaktige, men dette vil være tilfellet kun til en viss grad. Hvis f.eks. en persons høyde opprinnelig ble målt til nærmeste centimeter, kan det anonymiserte datasettet uttrykke en høyde som kan være +/- 10 cm. Hvis denne teknikken brukes effektivt, vil en tredjepart ikke kunne identifisere en person, reparere dataene eller på annen måte finne ut av, hvordan dataene har blitt endret.

Teknikken bør kombineres med andre anonymiseringsteknikker slik som fjerning av åpenlyse attributter og kvasiidentifikatorer. Støynivået bør avhenge av det nødvendige informasjonsnivået og konsekvensene for personers privatliv, hvis de beskyttede attributtene blir offentliggjort.

**Hvilken garanti har man for at teknikken virker**

- *Utskillelse:* Det er fortsatt mulig å skille ut oppføringer relatert til en person (eventuelt på en ikke-identifiserbar måte), selv om oppføringene er mindre pålitelige.
- *Sammenkobling:* Det er fortsatt mulig å koble sammen oppføringer relatert til den samme personen, men oppføringene er mindre pålitelige. Det betyr at en reell oppføring kan kobles sammen med en falsk oppføring («støy»). I noen tilfeller kan en feil attributt utsette en registrert for betydelig og enda større risiko enn en korrekt attributt.
- *Uttrekking:* Uttrekkingsangrep kan la seg gjøre, men det vil være mindre sjanse for at de lykkes og falske positiver (og falske negativer) er mulige.

**Vanlige feil**

- *Å tilføre inkonsekvent støy:* Dersom støyen ikke er semantisk korrekt (dvs. den er uproporsjonert og ikke følger logikken mellom attributtene i et datasett), vil en angriper med tilgang til databasen kunne filtrere bort støyen og i visse tilfeller regenerere de opplysningene som mangler. Dersom datasettet er for spredt, kan det fortsatt være mulig å sammenkoble «støyopplysningene» til en ekstern kilde.
- *Dersom vi antar at støyen er tilstrekkelig:* Denne teknikken er et tillegg, som gjør det vanskeligere for en

angriper å finne personopplysningene. Med mindre støyen er høyere enn opplysningene i datasettet, skal man ikke anta at tilførsel av støy utgjør en selvstendig løsning for anonymisering.

### Ulemper ved tilførsel av støy

Et velkjent eksperiment med reidentifisering ble utført på kundedatabasen på Netflix, som omtalt i kapittel tre. Virksomheten offentliggjorde databasen etter at alle opplysninger som kunne identifisere kundene var blitt fjernet med unntak av anmeldelser og dato. Det ble tilført støy ved å øke eller senke anmeldelsene litt. På tross av dette fant man ut at 99 prosent av brukerne kunne unikt identifiseres i datasettet med åtte anmeldelser og datoer med 14 dagers forskyvning som utvelgelseskriterier. Dersom utvelgelseskriteriene ble senket (to karakterer og tre dagers forskyvning), kunne fortsatt 68 prosent av brukerne identifiseres.

### Permutasjon

Denne teknikken består av å bytte om på attributtenes verdier i en tabell, slik at noen av dem kunstig kobles sammen med andre registrerte. Dette er hensiktsmessig når det er viktig å bevare den nøyaktige fordelingen av de enkelte attributtene i datasettet.

Permutasjon er en spesiell måte å tilføre støy på. I en klassisk støyteknikk endres attributter med tilfeldige verdier. Det kan være vanskelig å generere konsekvent støy, og det er antakelig ikke tilstrekkelig å endre attributtverdiene litt for å gi godt personvern. Som alternativ endrer permutasjonsteknikkene verdiene i datasettet gjennom kun å bytte dem fra en oppføring til en annen. Denne ombyttingen betyr at verdienes intervall og fordeling vil være de samme, men at korrelasjonene mellom verdiene og personene forandres. Hvis to eller flere attributter har en logisk relasjon eller en statistisk korrelasjon, vil denne relasjonen ødelegges dersom de permuteres uavhengig av hverandre. Det kan derfor være viktig å permutere et sett relaterte attributter for ikke å bryte den logiske relasjonen, ellers vil en angriper kunne identifisere de permuterte attributtene og snu permutasjonen.

Hvis vi for eksempel har en undergruppe av attributter i et medisinsk datasett, «årsaker til innleggelse/symptomer/ansvarlig avdeling», finnes det i de fleste tilfeller en sterk logisk relasjon mellom verdiene, og om bare en av verdiene permuteres vil dette bli oppdaget, og det kan til og med reverseres.

Akkurat som tilførsel av støy, er det ikke sikkert at permutasjon sikrer anonymisering i seg selv, og den bør alltid kombineres med fjerning av åpenbare attributter/kvasiidentifikatorer.

### Hvilken garanti har man for at teknikken virker

- *Utskillelse:* Akkurat som ved tilførsel av støy, er det fortsatt mulig å utskille en enkeltpersons oppføringer, men de er mindre pålitelige.
- *Sammenkobling:* Dersom permutasjon påvirker attributter og kvasiidentifikatorer, kan det forhindre «korrekt» sammenkobling av attributter, både internt og eksternt, men tillater fortsatt «feilaktig» sammenkobling, ettersom en ekte opplysning kan settes på en annen registrert.
- *Uttrekking:* Det er fortsatt mulig å trekke ut opplysninger fra datasettet, særlig om attributtene er korrelert eller har sterke logiske relasjoner. Ettersom angriperen ikke vet hvilke attributter som har blitt permutert, må han/hun overveie om uttrekking er basert på en feilaktig hypotese og at det derfor bare er mulig med probabilistisk uttrekking.

### Vanlige feil

- *Å velge feil attributt:* Å permutere de minst sensitive attributtene eller attributter med lavest risiko, vil ikke i vesentlig grad øke beskyttelsen av personopplysningene. Hvis de sensitive/risikable attributtene fortsatt er koblet til det opprinnelige attributtet, kan en angriper fortsatt trekke ut sensitiv informasjon om enkeltpersoner.
- *Å permutere attributter tilfeldig:* Hvis to attributter har en sterk korrelasjon, vil tilfeldig permutasjon ikke gi noen sterke garantier. Denne vanlige feilen er illustrert i tabell 1.
- *Å anta at permutasjon er tilstrekkelig:* Akkurat som tilførsel av støy, gir ikke permutasjon anonymitet i seg selv og bør kombineres med andre teknikker som f.eks. å fjerne en åpenbar attributt.

### Ulemper ved permutasjon

Dette eksemplet viser hvordan tilfeldig permutasjon av attributter fører til dårlig personvern når det finnes logiske relasjoner mellom ulike attributter. På tross av at man har prøvd å anonymisere opplysningene, er det en lett sak å finne ut hva den enkeltes inntekt har vært avhengig av stilling (og fødselsår). Gjennom en direkte gjennomgang av opplysningene kan man for eksempel anta at administrerende direktør antakelig er født i 1957 og har den høyeste inntekten, mens den arbeidsledige er født i 1964 og har den laveste inntekten.

| År | Kjønn | Stilling | Inntekt (permutert) |
|---|---|---|---|
| 1957 | M | Ingeniør | 70000 |
| 1957 | M | Adm.dir. | 5000 |
| 1957 | M | Arbeidsledig | 43000 |
| 1964 | M | Ingeniør | 100000 |
| 1964 | M | Leder | 45000 |

Tabell 1. Eksempel på anonymisering ved hjelp avpermutasjon av korrelerte attributter som ikke virker.

**«Differential privacy»**

Differential privacy hører til gruppen randomiseringsteknikker (se Dwork, C.: «Differential privacy». I *Automata, languages and programming*, 2006*)* Men med en annen vinkling: Tilførsel av støy skjer før et datasett skal tilgjengeliggjøres. Differential privacy kan derimot brukes når den behandlingsansvarlige genererer anonymiserte visninger av et datasett og samtidig beholder en kopi av de opprinnelige dataene. Slike anonymiserte visninger blir normalt generert av en delmengde av spørsmål for en bestemt tredjepart. Delmengden inneholder randomisert støy som bevisst er tilført i etterkant. Differential privacy forteller den behandlingsansvarlige hvor mye støy som må tilføres, og i hvilken form, for å få nødvendige garantier for personvernet (se Ed Feltens *Protecting privacy by adding noise*, 2012). I denne sammenhengen er det spesielt viktig at man fortløpende overvåker (minimum for hver nye spørring) om det finnes mulighet for å identifisere en enkeltperson i spørreresultatet. Det må imidlertid være helt klart at teknikker for differential privacy ikke endrer de opprinnelige opplysningene, hvilket betyr at så lenge de opprinnelige opplysningene finnes, vil den behandlingsansvarlige kunne identifisere enkeltpersoner i resultatene av differential privacy-spørringene når man tar hensyn til alle sannsynlige og tenkelige hjelpemidler som kan komme til å bli brukt. Slike resultater må altså ansees for å være personopplysninger.

En fordel med en fremgangsmåte som bygger på differential privacy er at datasett gjøres tilgjengelige for en betrodd tredjepart som svar på en bestemt spørring, i stedet for at et enkelt datasett gjøres tilgjengelig. Til hjelp ved revisjon kan den behandlingsansvarlige føre en liste over alle spørringer og forespørsler for å forsikre at tredjeparter ikke får tilgang til opplysninger som de ikke har autorisert tilgang til. En spørring kan også anonymiseres, herunder ved hjelp av tilførsel av støy eller substitusjon for ytterligere å ivareta personvernet. Det er fortsatt et uløst forskningsproblem å finne en fungerende søkemekanisme som klarer å besvare spørringer noenlunde korrekt (det vil si med mindre støy), samtidig som personvernet ivaretas.

For å begrense uttrekking- og sammenkoblingsangrep, er det nødvendig å overvåke de spørringene som sendes fra en enhet og observere de opplysningene som fremkommer om de registrerte. Databaser med differential privacy bør derfor ikke brukes av åpne tilgjengelige søkemotorer som ikke har mulighet for å spore enhetene som gjør spørringer.

**Hvilken garanti har man for at teknikken virker**

- *Utskillelse:* Hvis det kun er statistikk som produseres og de valgte reglene er hensiktsmessige, skal det ikke være mulig å bruke svarene for å skille ut en enkeltperson.
- *Sammenkobling:* Ved å bruke flere forespørsler kan det være mulig å sammenkoble opplysninger om en enkeltperson fra to svar.
- *Uttrekking*: Det er mulig å trekke ut opplysninger om enkeltpersoner eller grupper ved hjelp av flere forespørsler.

**Vanlige feil**

- *Å tilføre for lite støy:* For å unngå sammenkobling med bakgrunnskunnskap er utfordringen å ha så få spor som mulig om hvorvidt en bestemt registrert eller gruppe av registrerte er i datasettet. Den største utfordringen for beskyttelse av personopplysninger, er å kunne generere tilstrekkelig med støy som legges til de opprinnelige resultatene, slik at den enkeltes personvern beskyttes samtidig som de gjenværende resultatene fortsatt er brukbare.

**Ulemper ved differential privacy**

*Å behandle hver spørring isolert:* En kombinasjon av søkeresultat kan gjøre det mulig å avsløre informasjon som skulle være hemmelig. Dersom spørrehistorikken ikke er lagret, kan en angriper konstruere flere spørringer mot en database med differential privacy som gradvis innsnevrer resultatet helt til det med sikkerhet eller stor sannsynlighet dukker opp et visst kjennetegn for en enkelt registrert eller en gruppe registrerte. Videre er det viktig å unngå å gjøre den feilen at man tenker at opplysningene er anonyme for en tredjepart, mens den behandlingsansvarlige fortsatt kan identifisere den registrerte i den opprinnelige databasen.

## Generalisering

Generalisering er den andre gruppen av anonymiseringsteknikker. Denne fremgangsmåten består i å generalisere, eller vanne ut, de registrertes attributter gjennom å endre på den relative størrelsesorden (for

eksempel et fylke i stedet for en kommune, en måned i stedet for en uke). Generalisering kan være effektivt for å forhindre å skille ut personer, men anonymiseringen fungerer ikke bestandig. Det kreves spesifikke og avanserte kvantitative metoder for å forhindre sammenkobling og uttrekking.

**Aggregering og k-anonymitet**

Aggregerings- og *k*-anonymitetsteknikker har til formål å forhindre at registrerte blir skilt ut gjennom å gruppere dem med minst k andre personer. For å oppnå dette generaliseres attributtverdiene slik at hver enkelt person har samme verdi. Gjennom for eksempel å senke detaljnivået (granulariteten) for et sted fra en by til et land kan et større antall registrerte inkluderes. Enkelte fødselsdatoer kan generaliseres til et intervall av datoer eller grupperes på måneder eller år. Andre numeriske attributter (for eksempel lønn, vekt, lengde eller doser av medisiner) kan generaliseres gjennom intervallverdier (for eksempel lønn 250 000 – 300 000 kroner). Disse teknikkene kan brukes når korrelasjonen av attributters nøyaktige verdier kan skape kvasiidentifikatorer.

**Hvilken garanti har man for at teknikken virker**

- *Utskillelse:* Ettersom samme attributt nå deles av k-brukere, skal det ikke lenger være mulig å utskille enkeltpersoner innen en gruppe av k-brukere.
- *Sammenkobling:* Muligheten for sammenkobling er begrenset, men det er fortsatt mulig å sammenkoble oppføringer innenfor grupper av k-brukere. Innenfor denne gruppen er sannsynligheten for at to oppføringer tilsvarer samme pseudoidentifikatorer 1/k (hvilket kan være betydelig høyere enn sannsynligheten for at slike opplysninger ikke kan sammenkobles).
- *Uttrekking:* Den største ulempen ved k-anonymitetsmodellen er at den ikke forhindrer noen form for uttrekkingsangrep. Hvis alle k-personer inngår i samme gruppe og det er kjent hvilken gruppe en enkeltperson tilhører, er det lett å finne frem verdien av denne egenskapen.

**Vanlige feil**

- *Å utelate noen kvasiidentifikatorer:* Et kritisk parameter i forbindelse med *k*-anonymitet er terskelen for k. Jo høyere verdien av k er, jo bedre blir personvernet. En vanlig feil er å kunstig øke verdien av k gjennom å redusere de brukte kvasiidentifikatorene. Å redusere kvasiidentifikatorer gjør det enklere å bygge opp klynger av k-brukere på grunn av andre attributters innebygde muligheter for identifisering. Dette er særlig et problem om en del av de er sensitive eller har en veldig høy entropi, slik som er tilfellet for veldig sjeldne attributter. Det er en alvorlig feil å ikke vurdere alle kvasiidentifikatorer når man skal velge hvilken attributt som skal generaliseres. Hvis noen attributter kan brukes for å skille ut en person i en klynge av k, har ikke generaliseringen beskyttet noen personer (se eksempelet i tabell 2).
- *Lav verdi av k:* Det fører til et liknende problem om man tildeler k en lav verdi. Hvis k er for lav, vil de enkelte personene i en klynge ha for stor vekt, og det vil være større risiko for at uttrekkingsangrep vil lykkes. For eksempel hvis k=2, vil sannsynligheten for at de to personene deler samme egenskap være høyere enn for k>10.
- *Ikke gruppere enkeltpersoner med samme vekt:* Det kan også føre til problemer med gruppering av personer med ulike fordeling av attributter. Innflytelsen av en enkeltpersons oppføringer på et datasett vil variere: Noen kommer til å representere en stor andel av verdiene, mens andres bidrag forblir ganske ubetydelige. Det er derfor viktig å sørge for at k er tilstrekkelig høyt slik at ingen enkeltpersoner representerer en altfor stor andel av verdiene i en klynge.

**Ulemper ved k-anonymitet**

Den største ulempen ved k-anonymitet er at det ikke forhindrer uttrekkingsangrep. I følgende eksempel vet angriperen at en bestemt person er i datasettet og ble født i 1964. Han vet også at personen har hatt et hjerteinfarkt. Hvis vi vet at datasettet stammer fra en fransk organisasjon, så bor personene i Paris, ettersom de første tre sifre i postnumrene i Paris er 750*.

| År | Kjønn | Postnr | Diagnose |
|---|---|---|---|
| 1957 | M | 750* | Hjerteinfarkt |
| 1957 | M | 750* | Kolesterol |
| 1957 | M | 750* | Kolesterol |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |

**Tabell 2.** Et eksempel på dårlig utført k-anonymisering.

**L-diversitet/t-nærhet**

*L*-diversitet utvider *k*-anonymitet for å sikre at deterministiske uttrekkingsangrep ikke lenger er mulig. Dette skjer ved å sikre at hver enkelt attributt har minst *l* ulike verdier i hver ekvivalensklasse.

Et grunnleggende mål er å begrense forekomsten av ekvivalensklasser med lav variabilitet i attributtene, slik at en angriper med bakgrunnskunnskap om en bestemt registrert alltid sitter igjen med en betydelig usikkerhet.

L-diversitet er en nyttig metode for å beskytte data mot uttrekkingsangrep når attributtenes verdier er godt fordelt. Det må understrekes at denne teknikken ikke kan forhindre lekkasje av opplysninger, hvis attributtene i en partisjon er ujevnt fordelt eller tilhører et lite intervall av verdier eller semantiske betydninger. L-diversitet kan imidlertid utsettes for probabilistiske uttrekkingsangrep.

T-nærhet er en videreutvikling av l-diversitet, som har som mål å skape ekvivalensklasser, som likner den opprinnelige fordelingen av attributter i tabellen. Denne teknikken kan brukes når det er viktig at dataene endres så lite som mulig fra de opprinnelige dataene. Til det formålet settes en ytterligere begrensning på ekvivalensklassen, nemlig at det ikke bare skal finnes *l* ulike verdier innen hver ekvivalensklasse, men også at hver verdi representeres så mange ganger som kreves for å avspeile den opprinnelige fordeling av hvert enkelt attributt.

**Hvilken garanti har man for at teknikken virker**

- *Utskillelse:* Akkurat som ved k-anonymitet, kan l-diversitet og t-nærhet sikre at oppføringene om en enkeltperson ikke skilles ut av databasen.
- *Sammenkobling:* L-diversitet og t-nærhet er ikke bedre enn k-anonymitet når det gjelder sammenkobling. Problemet er det samme som med alle klynger: Sannsynligheten for at samme verdier tilhører samme registrerte er høyere enn 1/N (hvor N er antallet registrerte i databasen).
- *Uttrekking:* Den største forbedringen som l-diversitet og t-nærhet gir sammenliknet med k-anonymitet er at det ikke lenger er mulig å konfigurere et uttrekkingsangrep mot en database med l-diversitet og t-nærhet med 100 prosent sikkerhet.

**Vanlige feil**

- *Å beskytte sensitive attributtverdier ved å blande dem med andre følsomme attributter:* Det er ikke tilstrekkelig å ha to verdier for et attributt i en klynge for å beskytte personvernet. Fordelingen av sensitive verdier i hver klynge bør likne fordelingen av disse verdiene i hele populasjonen, eller i det minste være enhetlig innenfor hele klyngen.

**Ulemper ved l-diversitet**

I tabellen nedenfor er det utført l-diversitet for attributtet «Diagnose». Hvis man vet at en person som er født i 1964 er i denne tabellen er det med høy sannsynlighet fortsatt mulig å anta at han har hatt et hjerteinfarkt.

| År | Kjønn | Postnr | Diagnose |
|---|---|---|---|
| 1957 | M | 750* | Hjerteinfarkt |
| 1957 | M | 750* | Kolesterol |
| 1957 | M | 750* | Kolesterol |
| 1957 | M | 750* | Kolesterol |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Kolesterol |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |
| 1964 | M | 750* | Hjerteinfarkt |

**Tabell 3.** En l-diversitetstabell der verdiene for «diagnose» ikke er jevnt fordelt.

| Navn | Fødselsdato | Kjønn |
|---|---|---|
| Smith | 1964 | M |
| Rossi | 1964 | M |
| Dupont | 1964 | M |
| Jansen | 1964 | M |
| Garcia | 1964 | M |

**Tabell 4.** En angriper som vet at disse personene finnes i tabell 3, kan trekke ut at de har hatt hjerteinfarkt.

(Dette kapittelet er en oversetting av kapittel 3 og 4 fra *Opinion 05/2014 on Anonymisation Techniques* (pdf) fra Artikkel 29-gruppen (WP29))

## Pseudonymisering

Pseudonymisering består i å erstatte ett attributt (oftest et unikt attributt) i en oppføring med et annet. Det er derfor fortsatt sannsynlig at den fysiske personen kan identifiseres indirekte. Hvis pseudonymisering brukes alene, vil det ikke bli et anonymt datasett. Metoden diskuteres allikevel her på grunn av de mange misforståelser og feil omkring bruken av pseudonymisering.

Pseudonymisering gjør det vanskeligere å knytte datasettet til den registrertes opprinnelige identitet og er derfor et nyttig sikkerhetstiltak, men ikke en metode for anonymisering.

Resultatet av pseudonymisering kan være uavhengig av den opprinnelige verdien (slik som er tilfellet med et tilfeldig valgt tall som genereres av den behandlingsansvarlige eller et etternavn som den registrerte velger), eller kan avledes av et attributts ellet et sett av attributter, for eksempel en hashfunksjon eller et krypteringssystem.

De mest utbredte pseudonymiseringsteknikkene er følgende:

- *Kryptering med en hemmelig nøkkel*
  I dette tilfellet kan innehaveren av nøkkelen lett reidentifisere hver enkelt registrert ved å dekryptere datasettet, ettersom personopplysningene fortsatt finnes i datasettet, bare i kryptert form. Hvis vi antar at man bruker et moderne krypteringssystem, er det kun mulig å dekryptere hvis man har nøkkelen.

- *Hashfunksjon*
  Dette er en funksjon som gir utdata med en fast størrelse fra inndata av enhver størrelse (inndata kan være et attributt eller et sett av attributter), som ikke kan tilbakeføres. Det innebærer at risikoen for tilbakeføring, som ved kryptering, ikke lenger finnes. Om intervallet med inndataverdier for hashfunksjonen er kjent, kan verdiene kjøres gjennom hashfunksjonen på nytt for å finne frem til den korrekte verdien for en bestemt oppføring. Hvis et datasett for eksempel har blitt pseudonymt gjennom en hashfunksjon på fødselsnummer, kan disse enkelt utledes ved å hashe alle mulige inndataverdier og sammenlikne resultatet med verdiene i datasettet. Hashfunksjoner pleier å utformes for å være relativt raske å beregne og er utsatt for «brute force»-angrep.[1] Det kan også lages forhåndsutregnede (pre-computed) tabeller som gir mulighet for tilbakeføring av en stor mengde hashverdier.

- *Bruk av en saltet hashfunksjon*
  (hvor en tilfeldig verdi betegnet «salt» føyes til det attributtet som hashes) kan begrense sannsynligheten for at inndataverdien blir funnet, men det kan fortsatt være mulig at man med alle sannsynlig og tenkelige hjelpemidler kan beregne den opprinnelige attributtverdien som skjules bak resultatet av en saltet hashfunksjon.[2]

- *Enveiskodet hashfunksjon med lagret nøkkel*
  Dette er en spesiell hashfunksjon hvor man bruker en hemmelig nøkkel som ytterligere inndata (dette er forskjellig fra en saltet hashfunksjon ettersom saltet ikke pleier å være hemmelig). En behandlingsansvarlig kan gjenta funksjonen på attributtet ved å bruke den hemmelige nøkkelen. Dette kan ikke en angriper gjøre ettersom antallet muligheter som må testes er så stort at det vil være umulig å gjennomføre.

- *Deterministisk kryptering eller nøkkelbasert hashfunksjon med sletting av nøkkel*
  Denne teknikken kan likestilles med å velge et tilfeldig tall som pseudonym for hver attributt i databasen og siden slette sammenlikningstabellen. Denne løsningen gjør det mulig[3] å begrense risikoen for sammenkobling mellom personopplysningene i datasettet og opplysninger om den samme personen i et annet datasett der et annet pseudonym blir brukt. Med en moderne algoritme blir det ved beregning vanskeligere for en angriper å dekryptere eller gjenta funksjonen, ettersom det skulle kreve at hver tenkelig nøkkel må prøves, forutsatt at nøkkelen ikke er tilgjengelig.

- *Tokenisering*
  Denne teknikken brukes ofte i finanssektoren (men er ikke begrenset til den sektoren) for å erstatte kort-id-nummer med verdier som er mindre anvendelige for en angriper. Teknikken har blitt utviklet av de tidligere nevnte og pleier å bygge på mekanismer for enveiskryptering eller en indeksfunksjon som tildeler et følgetall eller et tilfeldig tall som ikke er matematisk beregnet fra de opprinnelige data.

---

[1] Slike angrep består i å prøve ut alle tenkbare inndata for å bygge sammenlikningstabeller.

[2] Særlig om attributtets type er kjent (navn, fødselsnummer, fødselsdato osv.). For å legge til beregningskrav kan man bruke en hashfunksjon for nøkkelavledning, hvor den beregnede verdien hashes flere ganger med et kort salt.

[3] Avhengig av øvrige attributter i datasettet og på sletting av de opprinnelige opplysningene.

**Hvilken garanti har man for at teknikken virker**

- *Utskillelse*
  Det er fortsatt mulig å skille ut en enkeltpersons oppføring ettersom personen fortsatt identifiseres med et unikt attributt som er resultatet av pseudonymiseringsfunksjonen (= det pseudonyme attributtet).
- *Sammenkobling*
  Sammenkobling er fortsatt enkelt mellom oppføringer som bruker samme pseudonyme attributt for å vise til en og samme person. Selv om ulike pseudonyme attributter brukes for samme registrerte person, kan sammenkobling fortsatt være mulig ved hjelp av andre attributter. Det er kun om ingen andre attributter i datasettet kan brukes for å identifisere den registrerte, og om hver sammenkobling mellom det opprinnelige attributtet og det pseudonyme attributtet er fjernet (inkludert sletting av de opprinnelige data), at det ikke vil finnes noen åpenbar kryssreferanse mellom to datasett som bruker ulike pseudonyme attributter.
- *Uttrekking*
  Det er mulig med uttrekkingsangrep mot den registrertes virkelige identitet innenfor datasettet eller innenfor flere ulike databaser som bruker samme pseudonyme attributt for en enkeltperson. Det er også mulig om pseudonymer er selvforklarende og ikke skjuler den registrertes opprinnelige identitet tilstrekkelig.

**Vanlige feil**

- *Å tro at et pseudonymt datasett er anonymt*
  Behandlingsansvarlige tror ofte at det er tilstrekkelig å ta bort eller erstatte ett eller flere attributter for å gjøre et datasett anonymt. Det finnes mange eksempler som har vist at dette ikke er tilfellet. Bare å endre ID forhindrer ikke at noen identifiserer en registrert dersom kvasiidentikatorer fortsatt finnes i datasettet, eller om noen av verdiene for andre attributter fortsatt kan identifisere en person. I mange tilfeller kan det være like lett å identifisere en person i et pseudonymt datasett som blant de opprinnelige opplysningene. Det må gjøres mer for at et datasett kan regnes for å være anonymisert, blant annet å slette og generalisere attributter, eller sletting av opprinnelige data eller i hvert fall å bringe dem opp til et sterkt aggregert nivå.

**Vanlige feil ved å bruke pseudonymisering som teknikk for å redusere risikoen for sammenkobling:**

- *Å bruke den samme nøkkelen i ulike databaser*
  For å eliminere muligheten for sammenkobling av ulike datasett, er man svært avhengig av at å bruke en nøkkelbasert algoritme, og av at en enkeltperson tilsvarer ulike pseudonyme attributt i ulike sammenhenger. For å begrense muligheten for sammenkobling, er det derfor viktig at man ikke bruker samme nøkkel i ulike databaser.
- *Å bruke forskjellige nøkler («roterende nøkler») til forskjellige brukere*
  Det kan være fristende å bruke forskjellige nøkler for forskjellige grupper av brukere og endre nøkkel for hver gang (for eksempel å bruke samme nøkkel for å registrere 10 oppføringer relatert til samme bruker). Hvis dette ikke utføres korrekt, kan det imidlertid skape mønstre som til dels begrenser de tenkte fordelene. Ved for eksempel å rotere nøkkelen etter bestemte regler for bestemte personer, blir det lettere å sammenkoble oppføringene som korresponderer til en gitt person. Dersom man fjerner en pseudonym oppføring i databasen samtidig som en ny oppføring kommer inn, kan det signalisere at begge oppføringene relaterer til samme fysiske person.
- *Å beholde nøkkelen*
  Om en hemmelig nøkkel lagres sammen med pseudonyme oppføringer, og oppføringene blir kompromittert, kan en angriper med enkelhet sammenkoble de pseudonyme oppføringene til deres opprinnelige attributt. Det samme gjelder hvis nøkkelen er lagret separat fra opplysningene, men ikke på en sikret måte.

**Ulemper ved pseudonymisering**

- *Helsebeskyttelse*

| 1. Navn, adresse, føds.dato | 2. Periode med særlig bistandsytelse | 3. BMI | 6. Referansenummer på kohortanalyse |
|---|---|---|---|
| | < 2 år | 15 | QA5FRD4 |
| | > 5 år | 14 | 2B48HFG |
| | < 2 år | 16 | RC3URPQ |
| | > 5 år | 18 | SD289K9 |
| | < 2 år | 20 | 5E1FL7Q |

Tabell 5. Et eksempel på pseudonymisering ved hashfunksjon (navn, adresse, fødselsdato) som lett kan tilbakeføres.

Et datasett er opprettet for å undersøke forholdet mellom en persons BMI og utbetaling av særlig bistandsytelse. Det opprinnelige datasettet inneholdt den registrertes navn, adresse og fødselsdato, men disse opplysningene er blitt slettet. Referansenummeret for kohortanalysen ble generert fra de slettede opplysningene av en hashfunksjon. Selv om navn, adresse og fødselsdato er slettet fra tabellen, er det lett å beregne referansenumrene hvis den

registrertes navn, adresse og fødselsdato, samt hashfunksjonen er kjent.

- *Sosiale nettverk*

Det har blitt påvist (se A. Narayanan og V. Shmatikov, «De-anonymizing social networks», i 30th IEEE Symposium on Security and Privacy, 2009) at sensitive opplysninger om bestemte personer kan ekstraheres fra grafer over sosiale nettverk, på tross av at disse data er pseudonyme. En leverandør av et sosialt nettverk antok feilaktig at pseudonymisering var tilstrekkelig robust for å forhindre identifisering etter å ha solgt data til andre virksomheter for markedsførings- og reklameformål. I stedet for virkelige navn, brukte leverandøren kallenavn, men dette var helt klart ikke tilstrekkelig for å anonymisere brukerprofilene, ettersom relasjonene mellom forskjellige personer er unike og kan brukes som identifikator.

- *Lokasjoner*

Forskere ved MIT har analysert et pseudonymt datasett bestående av 15 måneders koordinater for 1, 5 millioner menneskers bevegelser i tid og sted over et område med en radius på 100 km. De fant ut at 95 prosent av befolkningen kunne skilles ut med fire lokasjonspunkter, og at bare to punkter var nok til å skille ut mer enn 50 prosent av de registrerte (et av disse punktene er kjent og er høyst sannsynlig «hjem» eller «kontor») (se Y.-A. de Montjoye et al: «Unique in the Crowd: The privacy bounds of human mobility», Nature, nr 1376, 2013). Dette ga bare en begrenset mulighet for å beskytte personvernet, selv om personenes identiteter ble pseudonymisert ved å erstatte deres virkelige attributter med andre etiketter.

**Besøksadresse**:
Tollbugata 3, 0152 Oslo

**Postadresse:**
Postboks 8177 Dep., 0034
Oslo

postkasse@datatilsynet.no
Telefon: +47 22 39 69 00

**datatilsynet.no**
personvernbloggen.no