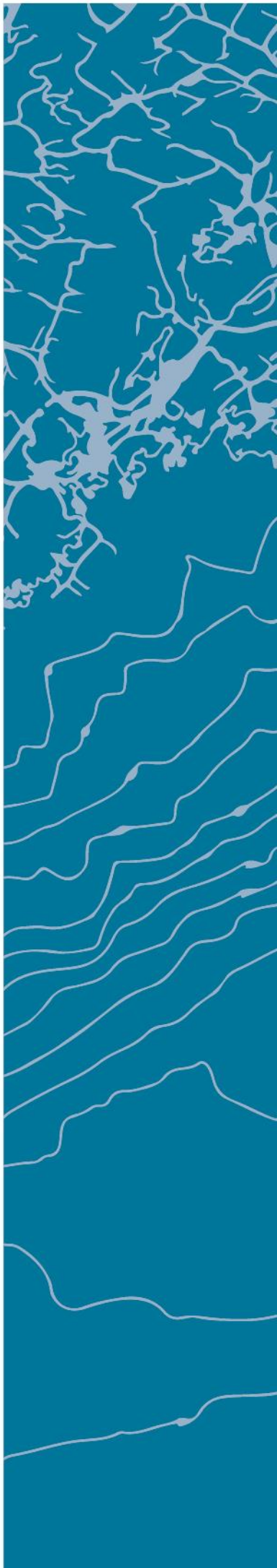
UNIVERSITETET I AGDER

Personvernutfordringer ved bruk av stordata i norsk offentlig sektor

SIGURD TVEIT KRISTOFFERSEN
FREDRIK SJULSTAD

VEILEDER
Dag Håkon Olsen

Universitetet i Agder, 2018
Fakultet for samfunnsvitenskap
Institutt for informasjonssystemer



Forord

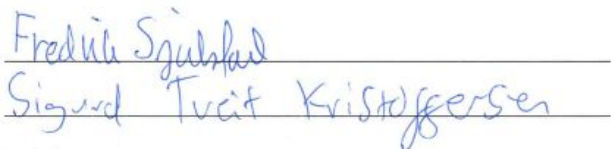
Denne masteroppgaven er et resultat av en studie utført i faget IS-501 Masteroppgave i Informasjonssystemer ved Universitetet i Agder, våren 2018. Studien er skrevet av Sigurd Tveit Kristoffersen og Fredrik Sjulstad, og avslutter deres mastergrad i Informasjonssystemer ved Universitetet i Agder. Studien tar for seg personvernutfordringer ved bruk av stordata i offentlig sektor, med intervjuer fra Skatteetaten, NAV, SSB og Datatilsynet. Vi vil takke alle informantene som bidro til denne studien.

Vi ønsker å gi en stor takk til vår veileder professor Dag Håkon Olsen ved Universitetet i Agder for god veiledning, innspill og nyttige samtaler.

Kristiansand, 03.06.2018.

Fredrik Sjulstad

Sigurd Tveit Kristoffersen



Sammendrag

Formålet med denne studien har vært å finne ut hvordan norsk offentlig sektor jobber med personvern tilknyttet stordata, og hvilke personvernutfordringer som oppstår. Vi har derfor utledet forskningsspørsmålet:

"Hvilke personvernutfordringer oppstår ved bruk av stordata i norsk offentlig sektor?"

Studien er gjennomført hos organisasjonene Skatteetaten, NAV, og SSB, der vi har fått innblikk i de ulike stordata-prosjektene de arbeider med. Vi har utført en kvalitativ case-studie med 14 semi-strukturerte intervjuer gjennomført høsten 2017 og våren 2018. Vi har også intervjuet Datatilsynet for å få innblikk hva de ser på som personvernutfordringer ved bruk av stordata.

Gjennom studien finner vi at den offentlige sektor ser store muligheter i bruk av stordata for effektivisering, statistikkformål og forbedring av tjenester. Prosjektene som utarbeides nå er fremdeles i startgropen, der det jobbes med å avdekke potensielle personvernutfordringer.

Vi ønsker å trekke frem noen sentrale utfordringer som går igjen hos organisasjonene, og blir sett på som spesielt viktige.

- Utilstrekkelige anonymiseringsteknikker som utgjør en fare for re-identifisering ved kobling av datasett.
- Sikring av tilgangskontroll og trygg lagring som sørger for at sensitiv informasjon beskyttes.
- Utfordringer knyttet til personvernprinsippene om dataminimering, formålsbegrensning og samtykke.
- Fare for bias og profilering ved feilaktige modeller og dårlig datakvalitet.

Offentlig sektor samler inn mye personsensitiv informasjon som utgjør en trussel for norske innbyggeres personvern hvis denne informasjonen misbrukes. Offentlige organisasjoner har ofte hjemmel til å samle inn flere personopplysninger enn det private organisasjoner har lov til. Oppgaven konkluderer med at det er viktig at offentlig sektor tar disse utfordringene på alvor, for å sikre tillit til oss som brukere av deres tjenester. Stordata bringer nye personvernutfordringer, og det er viktig at organisasjonene finner de beste metodene for å håndtere innsamling, lagring og analyse av stordata.

I praksis bidrar studien til å gi innsikt og bevisstgjøring om de personvernutfordringene offentlig sektor står overfor i sin videre utvikling av stordata-prosjekter. I teoretisk sammenheng kan studien bedre akademias forståelse av personvernutfordringer innenfor bruk av stordata i offentlig sektor.

Innholdsfortegnelse

Figurfortegnelse

Tabellfortegnelse

1. Innledning

Stordata har blitt et stadig viktigere fenomen, hvor organisasjoner tar i bruk stordata for å analysere behov og trender. En ser stadig at bedrifter tar i bruk sosiale medier som Twitter til å samle personers meninger om alt fra politiske initiativer til produktomtaler. Private organisasjoner har altså i lang tid sett fordelene ved bruken av stordata. Bedrifter kan analysere kundedata for å finne ut hvilke produkter man kan kjøpe inn mer av (Elgendy & Elragal, 2014).

I den norske offentlige sektoren blir stordata stadig mer sentralt. Kommunal- og moderniseringsdepartementet har i sin konferanse om stordata uttalt seg om at stordata er fremtiden for norsk offentlig sektor (Kommunal- og moderniseringsdepartementet, 2017). Det blir også gjort stadig større investeringer i norske stordata-prosjekter. For eksempel kjøpte Tolletaten nylig stordatatjenester for et større beløp fra en amerikansk leverandør for å kunne sortere ustrukturerte data som bilde og lyd (Datatilsynet, 2018).

I løpet av denne studien har vi sett at det ligger mange ubesvarte svar i bruken av stordata, spesielt med hensyn til personvern. Data som i utgangspunktet virker anonyme kan via avanserte teknikker bli brukt til å identifisere enkeltpersoner. En ser utfordringer ved at stadig større mengder med data blir samlet inn uten av at folk er klar over det (Datatilsynet, 2017). En finner også utfordringer i analyser ved stordata. Det uttrykkes bekymring om at organisasjoner tar i bruk stordata til å fatte stadig større avgjørelser basert på datasett av for liten kvalitet (Kaisler, Armour, Espinosa, & Money, 2013). Dersom enkeltpersoner blir en del av en feilberegning i en stordata-algoritme kan det være svært vanskelig å bevise at en er godartet (Maciejewski, 2016).

Datatilsynet sier selv at en av de største utfordringene ved stordata er anonymitet. Gjennom innsamling av et stadig større sett av datakilder er det en bekymring for at individer kan risikere å bli identifisert. De ser også at stordata legger press på generelle personvernprinsipper, og at individets rettigheter står på spill (Datatilsynet, 2017).

Den stadige større viktigheten av stordata i offentlig sektor, de ubesvarte personvernutfordringene og Datatilsynets bekymringer har vært med på å forme denne oppgaven. Vi har i denne studien ønsket å undersøke hvilke personvernutfordringer man finner i norsk offentlig sektor når det gjelder bruken av stordata.

Denne studien har hatt fokus på tre norske offentlige organisasjoner. Vi har studert stordata-prosjekter hos NAV, Skatteetaten og SSB. I tillegg har Datatilsynet blitt intervjuet, for å gi deres synspunkter som tilsynsorgan og ombud. Studien er en kvalitativ case studie, der det offentlige danner grunnlaget for caseoppgaven. Det er blitt utført totalt 14 intervjuer med ansatte i organisasjonene. Dette omfatter sentrale nøkkelpersoner i stordata-prosjekter. Studien ønsker å belyse hvilke personvernutfordringer en finner i norsk offentlig sektor. Vi har dermed formulert følgende forskningsspørsmål:

«Hvilke personvernutfordringer oppstår ved bruken av stordata i norsk offentlig sektor?»

1.1 Motivasjon

Bakgrunnen for denne studien faller først og fremst på vår interesse for konseptet stordata. Gjennom litteraturen ser vi at stordata stadig blir et viktigere fenomen både i den private og offentlige sektor. Begrepet i seg selv omfatter flere konsepter vi finner spennende, blant annet teknologier som maskinlæring, skylagring og stordataanalyse. Vi har hatt et ønske om å lære hvordan disse teknologiene fungerer i praksis og hvordan de påvirker menneskene rundt dem. Stordata var derfor et naturlig konsept å basere oppgaven på.

Videre fant vi i vår forstudie, utført høsten 2017, at stordata er på full vei inn i norsk offentlig sektor. Flere nyhetssaker og offentlige rapporter publisert av blant annet regjeringen, tyder på at dette vil være et satsningsområde fremover. Kommunal- og moderniseringsdepartementet sier selv et stordata er fremtiden for norsk offentlig sektor (Kommunal- og moderniseringsdepartementet, 2017). Flere store aktører, deriblant NAV, Skatteetaten, og Tolletaten er i gang med ambisiøse stordata-prosjekter (Datatilsynet, 2018). Å kunne få innsikt i hvordan norske offentlige organisasjoner anvender stordata i sine prosjekter ville være både interessant og lærerikt. Samtidig så vi at utgangspunktet i den norske offentlige sektoren ville kunne gi et bidrag utover de mangfoldige artiklene som en finner rundt konseptet stordata. Gjennom litteratursøk finner vi at stordata forskning med utgangspunkt i den offentlige sektor, stort sett baserer seg på USA.

Senere besluttet vi å avgrense oss til begrepet personvern. Nyhetsbildet er stadig preget av artikler som omfatter lekkasjer og datainnbrudd i større organisasjoner. I fjor preget Equifax-skandalen nyhetsbildet, hvor kundedata ble lekket (The New York Times, 2017). Nå nylig ble det satt et større fokus på personvern i bruken av sosiale medier da en fant ut at Cambridge Analytica hadde tatt i bruk store mengder med data fra Facebook-brukere uten samtykke (The New York Times, 2018). Videre har en også hatt utfordringer i norsk offentlig sektor i form av tjenesteutsetting til utlandet (NRK, 2017).

Desto mer aktuell blir studien med tanke på den nye personvernforordningen, GDPR, som trådte i kraft 25. Mai i år (van Loenen, Kulk, & Ploeger, 2016). Flere organisasjoner var denne våren i en omstillingsprosess, hvor applikasjoner og systemer var nødt til å følge det nye regulativet. Vi tenkte selv at det måtte være spennende å se hvordan slike personvernforordninger kan spille inn på stordata-prosjektene. Dette gjorde oss interesserte i å utforske stordata og personvern som et veldig aktuelt tema.

1.2 Avgrensninger

Masteroppgaven er avgrenset til NAV, SSB og Skatteetaten, som representerer norsk offentlig sektor. En forstudie ble også gjennomført høsten 2017 i forberedelse av denne masteroppgaven. Datatilsynet har også bidratt med synspunkter til studien. Studien er tidsbegrenset til et semester, som gjenspeiler omfanget.

1.3 Begrepsavklaringer

Tabell 1 - begrepsavklaringer

Begrep	**Definisjon**
Stordataanalyse	Stordataanalyse er prosessen om å utforske varierte og store datasett for å finne ukjente sammenhenger og mønstre som kan hjelpe organisasjoner med å ta bedre avgjørelser (Rouse, 2016).
GDPR	GDPR er et juridisk rammeverk innenfor den Europeiske Union som setter retningslinjer for prosessering og innsamling av personlig informasjon (Lord, 2017).
Web Scraping	Web scraping er en beskrivelse av metoder for å samle informasjon fra internett. Ofte blir dette gjort ved å simulere menneskelige handlinger på en nettside (Techopedia, 2018).
Datavarehus	Datavarehus er et sted der all data som organisasjonen samler inn. Datavarehus samler data fra forskjellige kilder for lagring og analyse, og styrer tilgang (Rouse, 2016).
Anonymisering	Anonymisering går ut på å fjerne identifiserende informasjon. Dette sørger for at den originale kilden ikke blir kjent (Merriam-Webster, 2018).
Pseudonymisering	Refererer til en prosess som reduserer risiko for identifisering ved å fjerne attributter direkte linket til en person (Mourby et al., 2018).
Hjemmel	Kravet til rettsgrunnlag for handling eller beslutning defineres som hjemmel. For at offentlige myndigheter skal iføre plikter til private parter, må det ligge hjemmel i lov, rettspraksis, forskrift eller annen rettskilde (Reusch, 2017).

Datamodellering	En prosess der man dokumenterer komplekse systemer slik at det blir enklere å forstå. Den dokumenterer hvordan data henger sammen og overføres (Rouse, 2016). I denne oppgaven bruker vi datamodellering som beskrivelse for hva som former stordataanalysene.
Metadata	Informasjon som beskriver annen informasjon, som regel ordnet i en søkbar database (Gjersdal, 2018).
Profilering	Å generalisere fra kjente tendenser og kjennetegn ved å samle inn informasjon om en person (Merriam-Webster, 2018).
Bias	Å ha urettmessige fordommer mot en person eller en gruppe (Oxforddictionaries 2018).
De-identfisering	Å endre på data slik at sensitiv informasjon ikke kan bli identifisert siden det har blitt endret fra sin originale form. Eksempler på dette er personidentifiserende informasjon som navn, e-post og personnummer (Techopedia, 2018).

2. Teori og tidligere forskning

2.1 Stordata

Stordata beskrives som en milepæl i informasjonsalderen, og som har stor innflytelse på samfunnet innenfor områder som klima, biologi, helse og vitenskap (Yu, 2016). Som et teknisk fenomen, finnes det flere meninger for hvordan man kan definere stordata. Stordata på et generelt grunnlag, betyr datasett som ikke kan oppfattes, anskaffes, administreres og behandles av tradisjonell maskinvare og programvare (Chen, Lao & Liui, 2014).

En ofte brukt definisjon som kom på banen så tidlig som i 2001, er 3V-modellen (Chen, Lao & Liui, 2014). De tre V'ene står for volume, variation og velocity - volum, variasjon og hastighet (Victor, Lopez, & Abawajy, 2016).

- Volum (volume) refererer til størrelsen på datasettene, som i en stordata-setting kan karakteriseres som enorme. Volum kjennetegnes også som en av hovedattributtene innenfor stordata, hvor størrelsen på datasettene kan kvantifiseres fra terabyte og petabyte, og stadig større.

- Variasjon (variation) vil si at dataenes kommer fra stadig varierende kilder. Stordata omfavner ikke kun strukturerte datakilder, men også ustrukturerte data, slik som datalogger, sosiale medier og lignende. Dette inkluderer også filtyper som bilder, video og lydfiler, kilder det er vanskelig å kategorisere.

- Hastighet (velocity) er også et viktig kjennetegn ved stordata. Hastighet refererer til hvor raskt data kan genereres og overføres, altså frekvensen. Flere forskere diskuterer at det å levere stordata i 'real-time' er kravet for en effektiv stordata-plattform i dag (Elgendy & Elragal, 2014).

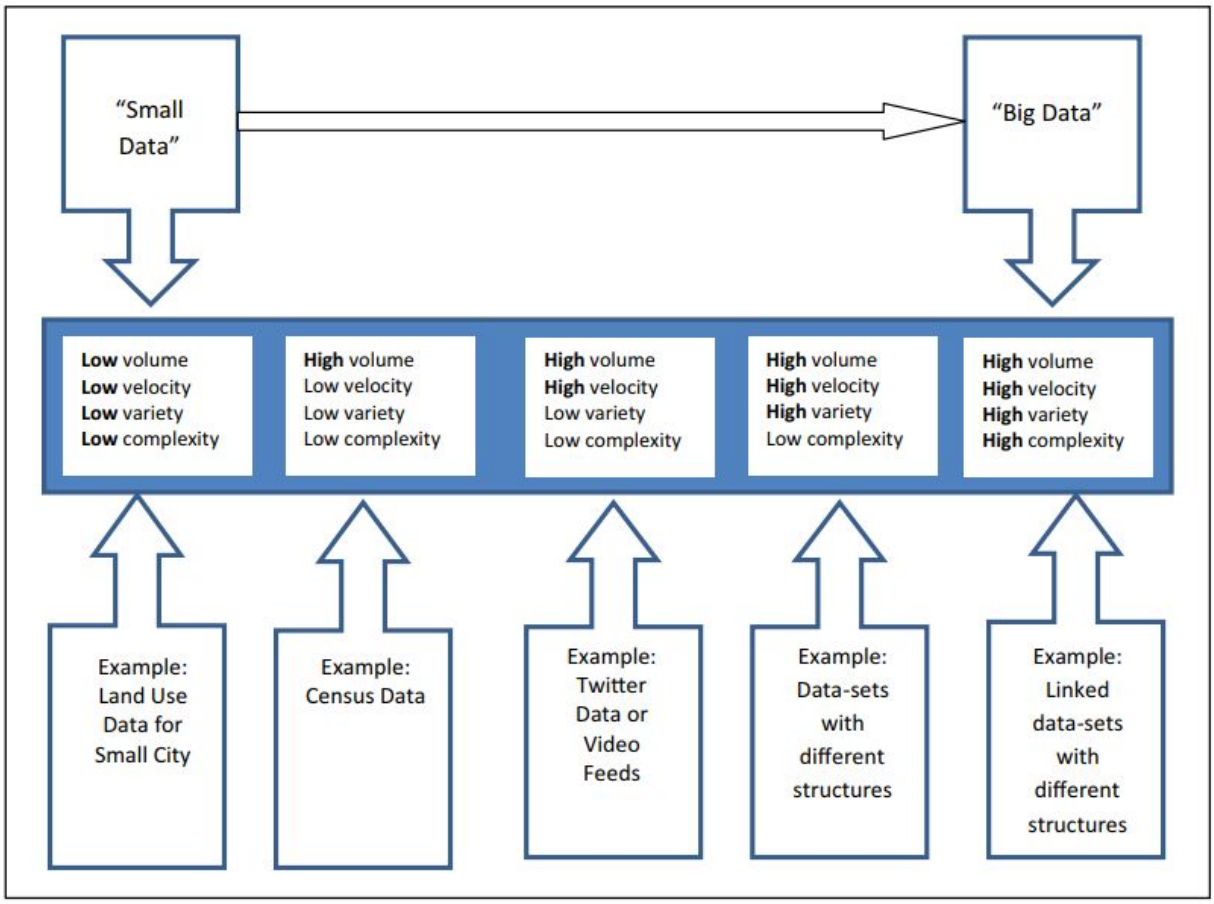


Fig 1. 3V-modellen. Volum, variasjon, hastighet (Desouza & Jacob, 2014).

Datakvalitet er også et viktig aspekt ved stordata. Datakvaliteten i store datasett kan være varierende, til det punktet hvor de er ufullstendige og blir ansett som tilnærminger av virkeligheten. Dette innebærer at dataene ikke korrekt representerer det man ønsker å samle inn data om, og gir et skjevt bilde av virkeligheten. En fjerde V, veracity (sannferdighet), som går på troverdighet til dataene, inkluderes derfor i modellen (Elgendy & Elragal, 2014).

I ettertid har forskere, blant annet IDC, en av de større aktørene når det gjelder stordata, utvidet denne definisjonen, ved å også inkludere verdi (value) i V-modellen. V-modellen poengterer nå at stordata går ut på å finne verdi bak datasettene, som ofte ellers er skjulte. En sentral utfordring ved stordata er å få verdi ut av store datasett, av ulike typer, strukturert og ustrukturert, i høyt tempo (Chen, Lao, Liui 2014).

2.2 Offentlig sektor og bruk av stordata

McKinsey Global Institute har kartlagt mulighetene som stordata kan gi i flere ulike kategorier innenfor offentlig sektor. Systemer kan automatiseres for å kalkulere risiko, samfunnsbehov kan

avdekkes, og ytelsen i ulike sektorer kan forbedres. Et eksempel er helsesektoren, der man kan analysere sykdomsmønstre, danne analytiske pasientprofiler og på generelt grunnlag forbedre folkehelsen (Sagioglu & Sinanc, 2013). Den Australske stat har definert i sin artikkel "Stordata strategi utfordringer" fire ulike potensielle muligheter for stordata i offentlig sektor. Smartere datastyring kan bidra til innsparinger. Personalisering av tjenester kan bedre finne behovene for individer og grupper av individer, og tilby mer spesifikke tjenester. Problemløsning og prediktiv analyse kan skape bedre beslutningsprosesser, og stordataanalyse kan bidra til innsparinger og effektivisering (Pandula Gamage, 2016). Stordata kan også bli brukt til å forsterke en organisasjons beslutningsevne ved å tilby større grad av synlighet om organisasjonens daglige operasjoner, samt forbedrede mekanismer for ytelsesmåling (McAfee & Brynjolfsson, 2012).

Det eksisterer utfordringer knyttet til stordata som er spesifikke for offentlig sektor. Fragmentering av funksjoner og kompetanse mellom ulike organisasjoner i offentlig sektor resulterer i ulike tolkninger av hva som kan deles med andre, og hva som ikke kan. Når offentlig sektor samler inn store mengder data, er det ofte innsamlingen fragmentert, der ulike avdelinger opererer i siloer når det kommer til deres egen informasjonsteknologi (Desouza & Jacob, 2014) Disse ulike tolkningene kan hindre samarbeid mellom de ulike organisasjonene som deltar i utveksling av informasjon. Ofte antar en stordata tilnærming at man har tilgang til all data som er nødvendig, noe som ikke er tilfellet i offentlig sektor (Malomo & Sena 2016). Det er ofte lite interoperabilitet mellom informasjonssystemene (Hilbert, 2016), som potensielt kan hindre integrasjonen av individuelle datasett inn i stordata (Desouza & Jacob, 2014).

Stordata blir også sett på som et sosio-teknologisk fenomen. På den ene siden ser man på stordata som et kraftig kraftig verktøy som kan løse ulike samfunnsproblemer som sykdommer, terrorisme og klimaendring. På den andre siden, er stordata også sett på som en trussel, en manifestasjon av "big brother", et verktøy som invaderer personvernet og reduserer friheter (Boyd & Crawford, 2012).

2.3 Personvern og stordata

Personvern
Personvern beskrives som en av de primære utfordringene innenfor databehandling (Pouloudi, 1999). Stordata brukes for å samle inn så mye data som mulig for å innhente kunnskap, noe som gjør at risiko for brudd på personvern er blant de største ulempene ved stordata (Soria-Comas & Domingo-Ferrer, 2015). Gjennom vårt litteratursøk finner vi at personvernutfordringer og stordata ofte er tett knyttet sammen.

Personvern er et relasjonelt og relativt konsept, der forskjellige verdier og forståelse av hva personvern innebærer vil være ulike for de som er involvert. (Pouloudi, 1999). Datatilsynet definerer personvern som retten til å bestemme over egne personopplysninger og retten til privatliv (Datatilsynet, 2016).

Grunnlovens paragraf 102 definerer personvern som «Enhver har rett til respekt for sitt privatliv og familieliv, sitt hjem og sin kommunikasjon. Husransakelse må ikke finne sted, unntatt i

kriminelle tilfeller». I 2014 ble integritet lagt til for å styrke vernet: «Statens myndigheter skal sikre et vern om den personlige integritet» (Datatilsynet, 2016).

Om data kan beskrives som sensitive og om de innehar personopplysninger, vil til en viss grad handle om hvilke egenskaper dataene har og hvorvidt de kan brukes for å identifisere enkeltpersoner. For et sett med informasjon finnes det ulike attributter som vi kan putte i 4 ulike kategorier.

1) Eksplisitt identifikator, unike attributter som identifiserer en person direkte, som personnummer.
2) Quasi-identifikator, attributter som har potensiale til å re-identifisere individer når man har nok informasjon slått sammen annen informasjon. Eksempel på dette kan være alder, karriære og postkode, som hver for seg ikke er identifiserende, men blir det ved sammenslåing.
3) Sensitiv informasjon. Informasjon som kan være interessant for andre å få tak i.
4) Annen. Informasjon som ikke passer i de andre kategoriene (Yu, 2016).

Modell for stordata og personvern

Med stordata må vi tenke annerledes om personvern. Tidligere modeller av personvern har presentert to modeller; "surveillance model" eller overvåkningsmodell, og "capture model" eller innsamlingsmodell. Begge disse modellene former måten vi diskuterer og beskriver personvern og data. Overvåkningsmodellen og innsamlingsmodellen er begge fokusert på å ivareta personvern når det kommer til innsamling og overvåkning. "Datafication model" eller datafikasjon, antar derimot at innsamlingen har skjedd allerede. Personvernutfordringene kommer derimot fra den nye kunnskapen, innsiktene og realitetene som skapes basert på dataen som er samlet inn (Mai, 2016).

1.) Den passive overvåkningen og kjøp av data vil ikke bli påvirket av innsamlingen av data ifølge den tradisjonelle overvåkningsmodellen. Dette sørger for at innsamlet data er en rasjonell representasjon og en refleksjon av virkeligheten.
2.) Innsamlingsmodellen mener at teknologi påvirker og endrer sosiale aktiviteter. Dette sørger for at dataene blir påvirket via selve innsamlingen gjennom vår bruk av teknologi. De innsamlede dataene blir derfor en forenkling av virkeligheten.
3.) Den foreslåtte datafikasjonmodellen av personvern fokuserer på prosessering og analyse fremfor innsamling. Datainnsamling er ontologisk orientert; den representerer hvordan verden eksisterer og presenterer fakta, om mennesker, aktiviteter, steder, tider, andre mennesker osv. Dataprosessering og analyse er epistemologisk orientert, der fokuset ligger på realiteter og fakta som genereres via prosessering og analyse. Personvernutfordringene skapes dermed først og fremst ved prosessering og analyse. (Mai, 2016).

En konsekvens av datafikasjon er at det skapes algoritmer som fører til avgjørelser som blir tatt uten menneskelig innblanding. De strategiske fordelene for organisasjoner er åpenbare og økende, men implikasjonene for det større samfunnet og individene er mindre klart (Mai, 2016).

MacAffe og Brynjolfsson mener at vi genererer mye data uten å være klar over hva konsekvensene er, hvordan det blir brukt og av hvem (McAfee & Brynjolfsson, 2012).

Bruk av algoritmer og stordata kan føre til potensielle urettferdige sosiale konsekvenser mellom grupper av individer, men kan også bli dratt enda lenger når data blir brukt til mer enn å forutse trender i grupper, men til å forutse handlingene til en enkeltperson (Newell & Marabelli, 2015). At valg blir tatt basert på stordata kan ha konsekvenser i form av diskriminering (Newell & Marabelli, 2015).

Cloud Security Alliance beskriver i sin rapport fra April 2013, de største utfordringene knyttet til stordata og sikkerhets og personvernutfordringer, som de har plassert i en modell:

1. Dataanalyse og datautvinning som bevarer personvern. Å anonymisere data for analyse er ikke nok for å sikre personvern. Beste praksis involverer ofte en implementasjon av overvåkning av systemet og loggføring. Å legge til "støy" i resultatene til dataanalysen kan være en metode å bedre personvern. En utfordring er å linke sammen ulike anonymiserte data og samtidig sikre konsistens mellom dem.

2. Kryptografisk datasikkerhet. Det finnes to forskjellige måter å kontrollere synligheten til data. Den ene er tilgangskontroll, den andre er å sikre data ved bruk av kryptografi. Ved å kombinere disse metodene, danner man standarden for de fleste data infrastrukturene som opererer med sensitive data. Angrep mot krypterte data er ofte langt vanskeligere å gjennomføre, enn angrep som er basert på å skaffe tilgang.

3. Granulær tilgangskontroll, som sørger for at tilgang blir gitt med langt høyere presisjon. En utfordring vil være å sikre hvem som skal ha tilgang og på hvilket nivå en gitt oppgave trenger tilgang. Stordata inneholder ofte massive data, som kan være vanskelig å kontrollere på tilgangsnivå. Jo mer komplekse applikasjoner blir, jo enklere er det å lukke alt bort. Dette vil gi veldig få tilgang til mulighet for analyse. Granulær tilgangskontroll blir nødvendig i kompliserte sikkerhetsmiljø (Cloud Security Alliance, 2012).

Fig 2. Personvernutfordringer, (Cloud Security Alliance, 2012).

2.4 Personvernprinsipper

Det europeiske parlamentet, rådet og kommisjonen nådde en enighet om den nye "General Data Protection Regulation" mot slutten av 2015, forkortet GDPR. Reguleringen sikrer rettighetene til beskyttelse av personlig data, og ble gjeldende fra sommeren 2018. GDPR ekspanderer reguleringer til data relatert til avansert dataanalyse. GDPR går lengre enn det som tidligere har vært regulert og kontrollert av data som inneholder personopplysninger. Dette innebærer at personlig data må innsamles for lovlige, tydelige og spesifiserte formål; personlig data må være relevante, tilstrekkelige og limiterte for å gjennomføre det formålet de ble innsamlet for. Personlig data må ikke kunne identifisere individer lengre enn det som er nødvendig for formålet med prosesseringen av dataen (van Loenen, Kulk, & Ploeger, 2016).

Et spørsmål er om GDPR utvider definisjonen av personlig data, der også pseudonymiserte data vil bli regulert under GDPR. I artikkelen "Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK" argumenteres det for at pseudonymisering ikke nødvendigvis produserer anonymiserte data, men refererer til en prosess som reduserer risiko for identifisering ved å fjerne attributter direkte linket til en person. Dette innebærer at personlig data som har blitt pseudonymisert, fremdeles kan gå inn under GDPR så lenge det er en fare for re-identifisering. For at personlig data ikke lengre skal gå inn under GDPR må det være anonymisert, og re-identifisering være umulig (Mourby et al., 2018).

Datatilsynet har opprettet en veileder som beskriver grunnleggende personvernprinsipper etter det nye regelverket GDPR som vi baserer oss på, da dette gir innblikk i hvilke personvernprinsipper de ser på som mest sentrale. GDPR i seg selv inneholder langt flere prinsipper, som blir for mange og detaljerte til å ha med i denne oppgaven. Disse prinsippene må følges for alle som behandler personopplysninger. Veilederen beskrives som en utredning av personvernprinsippene beskrevet i personvernforordningens artikkel 5, 6 og 9 (Datatilsynet, 2017).

Tabell 2 - Oversikt personvernprinsipper

Lovlig, rettferdig og gjennomsiktig	Et rettslig grunnlag må finnes for den behandlingen en virksomhet ønsker å utføre. For å oppnå dette må det innføres tiltak for å hindre at behandlingen ikke gå ut over det rettslige grunnlaget. Det må kontrolleres hva det er gitt samtykke til, og settes tekniske sperrer for all behandling som går ut over dette. Personopplysninger må behandles rettferdig. De må behandles i respekt for hva den som har utlevert personopplysninger har av interesser og rimelige forventninger. Det innebærer at behandlingen skal være forståelig og gjennomsiktig, ikke manipulerende eller fordekte. Automatiserte avgjørelser må bli tatt på forutsigbare og forståelige måter. Der er dermed viktig med åpenhet om opplysninger som kan føre til at enkeltpersoner blir delt ut i ulike kategorier og profilert. Opplysninger skal behandles gjennomsiktig. Dette betyr at behandlingen skal være oversiktlig og forutsigbar for den som har registrert opplysninger. Dette kan innebære personvernerklæring på virksomhetens nettsider om personvernregler (Datatilsynet, 2017).
Formålsbegrensning	Virksomheter må forsikre seg om at analyser stemmer med det originale innsamlings formålet. Man må sørge for at data som tidligere er samlet inn ikke brukes på nye måter som ikke var formålet med innsamlingen. Hvis opplysninger skal gjenbrukes må det innhentes nytt samtykke, eller det må være lovfestet (Datatilsynet, 2017).
Dataminimering	Det ligger store verdier i fremtidig bruk av data, som kan påvirke en virksomhets motivasjon og ønske om å beholde data. Både offentlige og private virksomheter kan ønske å beholde data for senere bruk for å kombinere med andre datasett, og dermed spare penger og skaffe seg nye innsikter. Datatilsynet skriver at prinsippet om dataminimering innebærer at man skal "begrense mengden innsamlede personopplysninger til det som er nødvendig for å realisere innsamlingsformålet". Det skal ikke hentes inn opplysninger om flere personer enn nødvendig eller som det er behov for (Datatilsynet, 2017).

Lagringsbegrensning	Prinsippet om lagringsbegrensning betyr at personopplysninger ikke skal lagres lengre enn det er nødvendig for formålet de er innhentet for, eller eventuelt anonymiseres (Datatilsynet, 2017).
Integritet og fortrolighet	Personopplysninger skal behandles slik at fortrolighet og integritet beskyttes. Metoder for å sikre dette er blant annet å benytte seg av tilgangskontroll og sikre seg mot uautorisert utlevering av personopplysninger. Videre skal opplysningene være tilgjengelige for autoriserte personer ved behov. Det må være mulighet for å spore endringer som gjøres i systemet og sikre at systemene beskytter mot sårbarheter, angrep og uhell (Datatilsynet, 2017).
Riktighet	Personopplysninger skal være korrekte, og hvis nødvendig skal de oppdateres. Et tiltak er å be brukere oppdatere sine opplysninger ved jevne mellomrom. Opplysninger som ikke er riktige skal slettes (Datatilsynet, 2017).
Ansvarlighet	Ansvarlighet er et prinsipp som ber om at man viser at man tar ansvar for korrekt behandling av personopplysninger. Proaktivt skal man etablere alle organisatoriske og tekniske tiltak for å sikre at regelverket følges (Datatilsynet, 2017).

Det skapes utfordringer ved å balansere personvern og utnyttelsen av stordata. Å samle korrekt informasjon om enkeltindivider, og samtidig ha god datakvalitet, vil ha følger for personvern ved prosessering av data (Maciejewski, 2016).

2.5 Stordata - livssyklus

For å få bedre innsikt i hvordan personlige data håndteres gjennom bruk av stordata, ønsket vi å finne en modell som tok for seg ulike faser for stordata og hvordan data behandles. Det er ingen fast modell som definerer livssyklusen til stordata, men heller flere ulike versjoner, noen er mer detaljerte og inndelte enn andre.Valget vårt falt på en modell som tar utgangspunkt i artikler fra Mehmood et al., (2016) og Chen et al. (2014). Modellen beskriver oversiktlig fasene innsamling og datagenerering, lagring og prosessering.

Innsamling og datagenerering

Det første steget i modellen er generering og innsamling av data. Enkeltpersoner legger igjen store mengder av høyst personlig informasjon, som medfører nye forventninger til for hva som bør holdes privat (Schadt 2012). Internett genererer store mengder data relatert til folks liv, innhentet fra søkemotorer, foruminnlegg, blogger og sosiale medier (Mehmood et al., 2016). I

offentlig sektor genereres mye data for hver enkelt person, som skaper mye data for offentlig administrasjon (Philip Chen & Zhang 2014).

På et individuelt nivå er ikke disse dataene nyttig, det er først når en ser sammenhengen på et stordatanivå at informasjonen blir brukbar. Data kan kombineres på nye måter som tidligere var umulig, da kan informasjon som vaner, oppførsel og hobbyer identifiseres (Mehmood et al., 2016).

Videre genereres også komplekse og mangfoldige datasett i stor-skala, som kan brukes til å predikere mer nøyaktig, og skape bedre beslutningsprosesser i henhold til naturkatastrofer, epidemier og andre samfunnsproblemer (Ekbia, Mattioli et al. 2015). Dette er datakilder som blant annet inkluderer kikkstrømmer, sensorer, og video. Innenfor generering av data ser en at hovedkildene i dag i stor grad kommer fra et IoT-perspektiv (Internet of things), lokalisasjonsdata og data generert for vitenskapelige formål. Handels- og logistikkdata fra bedrifter er også viktige kilder i et stordataperspektiv (Mehmood et al., 2016).

I datainnsamlingsprosessen tar en i bruk spesielle teknikker for å utvinne rådata fra kilden. Chen et al. tar for seg ulike metoder som ofte er brukt i denne prosessen. Dette gjelder blant annet bruken av logging for å logge filer som blir automatisk generert av kildesystemet, men også bruken av sensorer som fanger opp lyd, vær og temperatur (Mehmood et al., 2016). Når dataene er innsamlet tar en i bruk effektive overføringsmekanismer for å sende den til et lagringssystem for videre analyse (Mehmood et al., 2016).

Lagring

Den stadige veksten av data fører til større og strengere krav når det gjelder lagring, og skaper nye utfordringer og krav knyttet til arkitektur og sikkerhet. Lagring i et stordataperspektiv refererer til lagring og styring av store datasett, mens man samtidig må sørge for pålitelighet og tilgjengelighet. Det er viktig at lagringsinfrastrukturen kan bidra med å gi en lagringstjeneste som er pålitelig, samtidig som den kan gi tilgang til et kraftig grensesnitt for analyse av store mengder med data (Chen et al., 2014).

I følge Mehmood et. al (2016) består et lagringssystem av to deler; maskinvare og dataadministrasjon. Maskinvare vil si bruken av informasjons- og kommunikasjonsteknologi til ulike oppgaver, som for eksempel distribuert lagring. Dataadministrasjon (data management) refererer til bruken av programvare for å administrere store datasett (Mehmood et al., 2016).

Lagring er et stadig viktigere konsept innenfor stordata, bedrifter tar i bruk kraftigere infrastruktur enn tidligere for å kunne være konkurransedyktige. Forskere argumenterer for at det derfor trengs mer forskning når det gjelder lagring av store datasett (Chen et al., 2014).
Det er ofte høy redundans i datasettene, spesielt ved bruk av sensorer. For å sørge for at lagringskapasiteten benyttes effektivt kan det tas i bruk komprimeringsteknikker for å løse dette (Chen et al., 2014).

Analyse

Det siste steget i modellen, analyse, refererer til teknikker som brukes for å analysere og skape intelligens ut i fra stordata (Gandomi & Haider, 2015). Teoretikere har som nevnt sett på stordata analyser i sanntid for å være fremtiden når det gjelder stordata (Elgendy & Elragal, 2014).

Chen et al. (2014) skiller mellom sanntidsanalyser og 'offline' analyser. Sanntidsanalyser blir hovedsakelig brukt innenfor e-commerce og finans (Mehmood et al., 2016), og har blitt en trend innenfor stordata (Kacfah Emani, Cullot, & Nicolle, 2015). Siden data innenfor disse feltene endrer seg kontinuerlig er det nødvendig at resultater av analysen returneres hurtig, og med liten forsinkelse. En av de mest brukte metodene innenfor sanntidsanalyse er parallellprosessering; klynger (clusters) som benytter tradisjonelle relasjonsdatabaser (Chen et al., 2014).

Offline-analyser er vanligvis brukt med programvare med høye krav til maskinvare og responstid, dette gjelder blant annet til formål som maskinlæring og statistiske analyser. En ofte brukt metode for offline analyse er Hadoop som blant annet effektiviserer datakonvertering, datainnsamling (Chen et al., 2014), samt prosessering av mindre datasett (Daniell, Morton, & Ríos Insua, 2015).

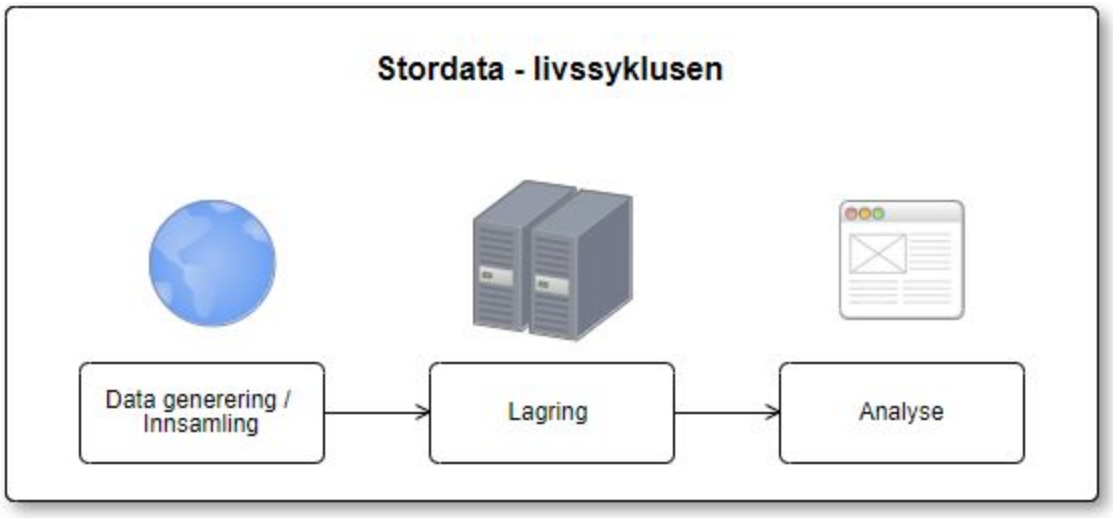


Fig.3 Livssyklusen til stordata (Mehmood et al., 2016; Chen et al., 2014).

2.6 Tidligere forskning - stordatalivssyklus

I denne delen presenterer vi potensielle pernsonvernutfordringer knyttet til innsamling, lagring og analyse vi har funnet gjennom litteratursøk. Dette gir et holistisk bilde av teoretiske utfordringer som har blitt forsket på tidligere, både i privat og offentlig sektor.

2.6.1 Innsamling

For stordata livssyklusen, er startpunktet innsamling av informasjon. Elektronisk kommunikasjon sørger for at vi legger fra oss data fra e-post, digitale meldinger og profiler. (Cumble, Church 2013).Ved datainnsamling er målet å sikre personvern ved å utvikle personvernmodeller og anonymiseringsteknikker for å finne potensielle risikoer for tap og angrep på personvern (Lei et al., 2014). For å kunne benytte data til å skape innsikter er man nødt til å vite hvor de kommer fra. Det er viktig å vite og dokumentere svakheter ved data som innsamles, for å forhindre utfordringer knyttet til bias og datakvalitet. Selv millioner av forskjellige sett med data blir ikke nødvendigvis representative (Boyd & Crawford, 2012).

Firmaer som tilbyr data bør bli undersøkt om hvilke metoder de har brukt for å respektere personvern ved innsamling av informasjon. Å dele informasjon med andre firmaer kan bryte med individets forventninger til personvern. Individer utgir informasjon innenfor et sett med personvernregler, så det er viktig å ha normer som styrer når, hvordan, hvorfor og hvor informasjon kan brukes (Hilbert, 2016).

Intern datainnsamling kan også by på utfordringer. For å maksimere innsiktene analyse vil det ofte kreve at data blir sammenslått med andre data i organisasjonen. Ved å kombinere på tvers av avdelinger skapes det unike utfordringer. I en undersøkelse av Desouza & Jacob (2014) blir dårlig styring av data nevnt som en utfordring for mange CIOs. Dette medfører at data ofte blir lokalisert innenfor enkelte avdelinger, og det å kunne slå sammen data blir en utfordring (Barocas & Nissenbaum, 2014).

Det er utfordringer knyttet til dataminimering ved innsamling. Å forklare alle mulige metoder data kan brukes på ved innsamling blir veldig vanskelig, fordi stordataanalyse er skapt for å hente ut gjemt informasjon, finne korrelasjoner mellom datasett og skape uforutsigbare slutninger. En følge av dette er at brukere trenger meningsfulle råd for å kunne ta valg når det kommer til bruken av deres data (Mantelero & Vaciago, 2015).

2.6.1.1 Samtykke

Data om et individ blir innsamlet, organisert og analysert for å danne et samlet bilde av en person. Et sentralt spørsmål er hvor mye av denne informasjonen du ønsker å beholde privat. Trenden i dag er å holde alt og anta at det vil bli benyttet til å skaffe innsikter over tid (Kaisler et al., 2013).

Ifølge artikkelen “Big data in public sector: lessons for practitioners and scholars" av Desouza et al. (2014), er det noen problemer med samtykke i forhold til stordata. Datasett uten sammenheng kan gjennom stordata-teknikker finne nye sammenhenger og korrelasjoner, som har ført til bekymringer knyttet til personvern (Desouza & Jacob, 2014). En etisk måte å drive en tjeneste på, vil være å spørre om tydelig samtykke, og kun for brukerens fordel (Hasan, Habegger, Brunie, Bennani, & Damiani, 2013). Utfordringer ligger i at en bruker ikke nødvendigvis forstår hva samtykke innebærer fordi de mangler teknisk kompetanse. I andre tilfeller blir det ikke gitt nok informasjon slik at en bruker kan tilstrekkelig gi samtykke (Perera, Ranjan, & Lizhe, 2015).

De færreste vil ta tid til å lese hva samtykket innebærer, hvilken data som innsamles, og hvordan den benyttes (Tan & Pivot, 2015).

En vesentlig verdi av demokratiske samfunn er at en borger har rett til informasjon om regjeringen, og hvordan avgjørelser og prosesser påvirker deres interesser. Dette innebærer en innsikt i hvordan potensielle algoritmer brukes for å hjelpe dem (Desouza & Jacob, 2014). Det eksisterer et problem med å bruke informasjon som er offentlig tilgjengelig, noe som muliggjør overvåking av personer som de kanskje ikke samtykker til. I noen tilfeller innebærer dette storstilt overvåkning og innsamling av data for å kontrollere og overvåke folkets holdninger (Liu & Yuan, 2015).

Artikkelen "Big Data: prospects and challenges", beskriver hvordan brukerpreferanser kan identifiseres i stordata-økosystemet, med data som tradisjonelt ikke krever samtykke. Kunden forstår ikke nødvendigvis hvordan data han gir bort kan brukes senere. Flere organisasjoner arbeider med store data, mange av dem uten kunnskap eller deltakelse av enkeltpersoner (Moorthy et al., 2015).

2.6.2 Lagring

Den stadige økning av data fører til at en må tenke nytt når det gjelder lagring. Når det gjelder tradisjonelle måter å håndtere data på brukes ofte relasjonsdatabase-behandlingssystemer (RDBMS). Problemet er imidlertid at denne formen for lagring kun støtter strukturerte data. Semi-strukturerte og ustrukturerte data er kjennetegn ved stordata, støttes ikke (Chen et. al, 2014). I tillegg kan det sies at det medfører svært mange kostnader å håndtere stordata ved bruk av tradisjonelle metoder for lagring (Leavitt, 2013). Infrastrukturen må stadig oppgraderes for å støtte mer lagringsplass (Leavitt, 2013). Dette har ført til at forskere har foreslått andre måter å håndtere lagring på deriblant bruken av skylagring. Fra et personvernperspektiv kan en argumentere for at fysisk lagring krever kompetanse. Blant annet for å hindre datainnbrudd, men også for at data går tapt. I tillegg må organisasjonen fokusere på rutiner for tilgangskontroll. Skylagring kan bidra å tette disse hullene, men har også sine begrensninger (Hashem et al., 2015).

Innenfor skylagring finne det utfordringer knyttet til blant annet tilgjengelighet, integritet, datakvalitet, lover og reguleringer. Når det gjelder tilgjengelighet er hovedutfordringen å kunne tilby en tilgjengelig plattform til enhver tid. Det settes krav til at data skal være tilgjengelig i løpet av kort tid (Hashem et al., 2015). Dette er utfordrende sett fra et personvernperspektiv da det også krever at plattformen er tilgjengelig i tilfeller ved datainnbrudd og hackerangrep (Zissis & Lekkas, 2012).

Når det gjelder stordatasikkerhet er integritet et viktig begrep. Dette vil i et skylagringsperspektiv stille krav til at data kun endres av autoriserte parter eller dataeiere slik at man kan forhindre misbruk. Videre er det viktig å sørge for at data som er lagret om brukere faktisk er korrekte, spesielt da brukere ofte ikke selv har tilgang til å se hva som er lagret om dem. Det foreslås å

utvikle mekanismer slik at brukere selv kan sjekke om data om dem vedlikeholdes (Hashem et al., 2015).

Datakvalitet innenfor skylagring har blitt et problem da data samles inn fra varierende kilder. Det er utfordrende å få data med høy kvalitet fra store og varierende datasett. I skyen skiller man høykvalitetsdata fra data med lavere kvalitet ved bruk av konsistens. Dersom data fra nye kilder er konsistente med andre kilder, så er dataene av høy kvalitet (Hashem et al., 2015). Lover og reguleringer er problematisk da ulike land har forskjellige lover når det gjelder datasikkerhet. Tankard et al trekker frem et eksempel hvor overvåkning av ansatte i flere land er ulovlig, men at elektronisk overvåking i spesielle tilfeller er tillatt (Tankard, 2012). Dette gjør at en kan stille seg spørsmål om lover gir tilstrekkelig med beskyttelse (Hashem et al., 2015).

2.6.2.1 Tilgangskontroll

Når det gjelder å beskytte personvernet er det ikke kun anonymisering som er et effektivt virkemiddel, tilgangsstyring er vel så viktig (Cavanillas, Curry, & Wahlster, 2016).

Tilgangskontroll gjennom autorisering er en av de vanligste måtene for å beskytte mot ulovlig bruk av innsamlet data. Passord har historisk sett vært den vanligste måten å sikre tilgangsstyring. Det finnes begrensninger for sikkerheten som passord tilbyr, ofte benyttes passord som er enkle å huske. Dette utgjør en risiko da personer ofte kan basere passordet sitt på detaljer om seg selv eller familiemedlemmer. Her kan to-faktor autentisering sørge for å øke sikkerheten fremfor kun bruk av passord (García Márquez & Lev, 2017, p. 83). Videre kan en ta i bruk rollebasert tilgangskontroll for å for å sørge at autoriserte brukere innenfor organisasjonen har tilgang. Å definere rollene, og å sette relevante tilgangsbegrensninger kan være utfordrende. Det er gjerne slik at pakker med verktøy som håndhever personvernregler er skreddersydd for ulike miljøer for å forenkle denne prosessen (Miller & Mork, 2013).

Zeng et al. foreslår tilgangsbegresninger basert på innhold. Dette fordi det ofte er det kan være både tidkrevende og vanskelig å sette tilgangsbegrensninger manuelt. På denne måten slipper en å identifisere hvert enkelt objekt. På den andre siden er ikke denne formen for tilgangsbegrensning like streng som rollebaserte tilganger, hvor den enkeltes behov blir identifisert (Zeng, Yang, & Luo, 2013).

Enkelte hevder også at de fleste organisasjoner som benytter stordata mangler tilstrekkelig med tilgangskontroll (Kshetri, 2014). Det er også slik at krypterte datasett kan trenge å bli dekryptert for analyse. Dersom organisasjoner tar i bruk ulike kryptografiske nøkler kan det bli en utfordring å dekryptere og rekryptere igjen. Risikoen for lekkasje kan øke og det kan være nødvendig med ytterligere datakraft (Mehmood et al, 2016).

Likevel er det slik at offentlige sjeldent deler data med tredjeparter da det ligger utfordringer knyttet til lover og reguleringer ved å ta i bruk ekstern arbeidskraft. En ønsker ikke at organisasjonens rykte kan risikere å bli ødelagt som følger av misbruk. Cavanillas et al. (2016)

referer til at de offentlige ønsker strengere reguleringer når det gjelder lagring og datatilgang, spesielt når det gjelder bruken av skytjenester (Canvillas et. al, 2016).

2.6.3 Analyse

2.6.3.1 Anonymisering og faren for re-identifisering

Robuste anonymiseringsteknikker kan benyttes for å hindre at personene i datasettene blir identifisert (García Márquez & Lev, 2017), og er en mulig løsning for å løse de konfliktene som oppstår mellom personvernprinsipper og stordata (Soria-Comas & Domingo-Ferrer, 2015). I stordatasammenheng kan man ikke garantere full anonymitet, og det eksisterer en risiko for at enkeltpersoner kan bli identifisert dersom datakilder settes sammen. Rådata om et individ kan kombineres med en adresse og kan være nok til å bli identifisert (García Márquez & Lev, 2017). Stadig mer kreative og sofistikerte metoder for re-identifisering skapes ettersom informasjon blir mer tilgjengelig (Daries et al., 2014).

Risiko for reidentifisering øker ved sammenkobling av datakilder, selv om dataene hver for seg er anonyme. Anonymisering som metode for å sikre personvern blir dermed mindre effektiv. En stor trussel for stordataanalyse vil være muligheten til å søke gjennom data for spesifikke enkeltindivider (Jensen, 2013). Ofte vil aggregerte data satt sammen fra forskjellige kilder for å skaffe innsikter som går ut over det en person er villig til å gi samtykke for. Det er en feiltagelse å anta at det å anonymisere data før det blir gitt ut til tredjeparter gjør det umulig å identifisere enkeltpersoner. Forskere har funnet forskjellige metoder og teknikker som kan brukes til re-identifisering, blant annet prediktive metoder som gjør anonymisering i enkelte tilfeller umulig (Kshetri, 2014). Det finnes verktøy som kan hjelpe med å måle risiko for re-identifisering, og sette opp parametere slik at risikoen kan bli minimert og utbytte maksimert. (Taneja, Kapil, & Singh, 2015). I offentlig sektor er de-identifisering av personlige data viktig for å sikre personvern, spesielt når man prøver å sette sammen små datasett der individer potensielt kan re-identifiseres De fleste organisasjoner de-identifiserer data før de deler dem, men etterhvert som volumet vokser øker risikoen for re-identifisering. Størrelsen på dataen og dens kompleksitet vil spille inn på hvor lett det er å forutsi faren for re-identifisering (Fola & Vania, 2016).

Det er mulig å bruke data aggregering til å konvertere semi-anonyme data eller ikke identifiserbar informasjon, til ikke-anonyme og potensielt identifiserbar informasjon (Kshetri, 2014). Data bør derfor anonymiseres tidlig i analyse-fasen for å hindre at personer kan identifiseres (García Márquez & Lev, 2017).

For en som analyserer stordata, er målet å få korrekte resultater, og samtidig holde sensitive data anonymisert. For å oppnå dette kan man velge å modifisere data før algoritmene for analyse blir implementert, og benytte seg av sikre protokoller for å sørge for at sikkerheten av sensitive og private data som er inne i de nye modellene som skapes (Lei et al., 2014).

2.6.3.2 Datakvalitet

For datakvalitet ligger det personvernutfordringer i å samle inn og prosessere informasjon om individer. En kan risikere at stordata modellene fører til diskriminering. Det kan også forekomme feil i datasettene som kan føre til at individer settes i bås. Innenfor stordata- analyse kan det forekomme menneskelige feil, datasett kan misforstås. Maciejewski skriver at dersom en faller innenfor det han referer til som "den digitale uregelmessighets-modellen", kan det være svært vanskelig å bevise at en ikke hører til denne kategorien (Maciejewski, 2016). Metoder for å verifisere datakvalitet er derfor en kritisk utfordring for stordata (Bertino, 2013).

Fra et analyseperspektiv er det en risiko at personer mistolker de konklusjonene stordata skaper. Data som mistolkes kan medføre konsekvenser for enkeltindivider. Det kan være en utfordring å bevise at noen har gjort noe galt gjennom analyse av stordata (Maciejewski, 2016).

Begrensninger og bias vil spille inn uansett hvor stor data man benytter seg av. Det er viktig å forstå og gjøre seg bevisst over dette, slik at man kan minimere sjansene for feiltolkning (Boyd & Crawford, 2012). Dersom beslutninger skal tas ut ifra modeller må informasjonen være nøyaktig. Organisasjoner må spørre seg selv om hvor mye data som er nok til å ta en beslutning, eller komme med et estimat (Kaisler, Armour, Espinosa, & Money, 2013).

Videre diskuterer Rogge, Agasisti & De Witte at organisasjoner som tar i bruk stordata bør være klar over at feil i datamodellering og analyser kan føre til store konsekvenser. I stordata-sammenheng er det ikke uvanlig med tilfeller av manglende data, feil i målinger, feil i duplikater, så vel som inkonsistens. Forfatterne tar utgangspunkt i en case hvor spørringer i søkemotoren til Google ble brukt for å overvåke influensasykdom. Prosjektet førte til en langvarig overestimering av influensa utbredelsen på bakgrunn av manglende kvalitet i datasettene (Rogge, Agasisti, & Witte, 2017).

Det er viktig å forstå egenskapene og begrensningene til dataene som samles inn. Selv med store data betyr ikke det nødvendigvis at dataene er representative og tilfeldige. Ofte sørger stordata for at mønstre som ikke er signifikante kommer frem, da korrelasjoner skapes som ikke nødvendigvis har noen relevant betydning. Et eksempel var aksjemarkedet og produksjon av smør i Bangladesh som hadde en sterk korrelasjon (Boyd et al., 2012).

2.6.3.3 Utbytte - balansegangen mellom brukbarhet og personvern

En annen utfordring knyttet til anonymisering gjelder blant annet utbytte. Dersom et datasett er for aggregert kan det være vanskelig å dra nytte av dataen. Mehmood et al. (2016) skriver at det er viktig å finne en balansegang mellom å tilby tilstrekkelig anonymitet og det å kunne få utbytte av stordata. Organisasjoner tar gjerne i bruk avanserte anonymiseringsteknikker for å fjerne identifiserbare felter fra et datasett, dette kan imidlertid føre til at dataen blir mindre brukbar (Mehmood et al., 2016).

Desouza & Jacob (2014) definerer denne problemstillingen innenfor den offentlige sektoren som balansegangen mellom individuelle rettigheter og offentlig interesse. De refererer til en forskningsgruppe som publiserte personidentifiserende informasjon om våpeneiere i New York i form av et visuelt kart. På den ene siden kan slik informasjon være for befolkningens beste, på den andre siden bryter det flere personvernprinsipper (Desouza & Jacob, 2014).

Forskere argumenter også for at det trengs mer forskning når det gjelder denne problemstillingen, og at det er et av de viktigste punktene når det gjelder stordata. Spesielt innenfor den offentlige sektor (Rogge et al., 2017).

3. Forskningstilnærming

I denne delen beskriver vi forskningsmetoden vi benyttet gjennom studien. Studien har blitt gjennomført som en kvalitativ case-studie, med temaet «personvernutfordringer i norsk offentlig sektor». En casestudie kan defineres som en «ting» som undersøkes; som et tema, et prosjekt, et informasjonssystem, en prosess eller lignende. Man ønsker å gå i dybden og skaffe detaljert innsikt, med de kompliserte prosessene og forhold som hører til (Oates, 2006). En god case-studie innebærer at den innehar

- Originalitet, at den er uvanlig og derfor av generell offentlig interesse.
- At problemene den utforsker er viktige for offentligheten, enten teoretisk, politisk eller praktisk.

Hvis case-studien treffer disse to punktene er den signifikant (Yin, 2009). Vårt valg av forskningsspørsmål var i stor grad basert på disse to punktene. Vi anser personvernutfordringer knyttet til stordata som av stor generell offentlig interesse, spesielt knyttet til offentlig sektor. Studien er original i den form at stordata-prosjektene i norsk offentlig sektor er et relativt nytt fenomen, som ikke er forsket mye på enda. At GDPR implementeres på samme tid som studien utføres, gjør at studiens funn blir relevante og originale.

Studiens analyse er basert på de semi-strukturerte intervjuene vi gjennomførte, med en kvalitativ metode. Kvalitativ forskning handler først og fremst å se på det store bildet og de mønstrene som finnes der, fremfor kvantitative studier som fokuserer på analyse av noen få faktorer gjennom bruk av tall og målinger (Hellevik, 2003). En utfordring med kvalitativ metode er at den er mer avhengig av dyktighet fra den som utfører studien. Det er derfor viktig å ha en strukturert tilnærming når man gjennomfører kvalitative studier (Oates, 2006).

3.1 Filosofisk paradigme

I bunnen av oppgaven ligger det et filosofisk paradigme, som antar at verden kan beskrives med delte antagelser og måter å tenke på (Oates, 2006). To ulike filosofiske paradigmer blir sett på som mest relevant; positivisme og fortolkende. Det positivistiske paradigme innebærer to antagelser om verden:

- At man kan studere verden og dens lover og mønstre fra et objektivt standpunkt, uavhengig av den som tolker det.
- At verden følger orden og regler, ting ikke skjer tilfeldig.

Positivisme er basert på den vitenskapelige metode, med hensikt om å finne felles lover, regler og mønstre (Oates, 2006). Teorier og forklaringer blir sett på som det beste vi har for øyeblikket, men man åpner for at disse i senere tid kan bli motbevist og endret (Oates, 2006).

I kontrast til positivisme finner vi fortolkende paradigme, som har en oppfatning om at verden eksisterer ut i fra flere subjektive sannheter. Virkeligheten blir skapt gjennom det individet som observerer omgivelsene. Man prøver å forstå fenomener ut i fra de verdiene og meningene mennesker gir dem, og ser på hvordan verden blir tolket individuelt. Målet er å gi en detaljert beskrivelse av ulike kontekster, og beskrive hvordan fortolkninger kan endres over tid og være forskjellige fra en gruppe eller en person (Creswell 2009; Oates, 2006).

Da stordata først og fremst er et teknisk fenomen, har vi benyttet oss av en positivistisk tilnærming til oppgaven. Personvernutfordringene som blir beskrevet er i stor grad knyttet til juridiske, organisatoriske og tekniske utfordringer ved stordata, og er dermed ofte felles for flere av organisasjonene, avhengig av hvor langt de har kommet med å avdekke disse utfordringene. Vi legger dermed til grunn at det finnes objektive sannheter om hvilke personvernutfordringer som finnes, selv om vi ser at ansatte til en viss grad vil subjektivt legge vekt på hva de definerer som utfordringer.

3.2 Utvalg av informanter

Gjennom forstudien utført i 2017 kom vi i kontakt med ansatte både i Skatteetaten og NAV. De samme personene var til hjelp med å finne nye informanter innenfor organisasjonen ved gjennomføringen av masterstudiet. Informantene var dermed valgt ut ved hjelp av ikke-sannsynlighetsutvalg, altså ikke ved tilfeldig utvalg (Oates 2006, Hellevik 2003).

Som metode for å oppnå dette, brukte vi "snøballutvalg", hvor vi fikk kontakt med organisasjonen gjennom noen nøkkelpersoner som kunne dirigere oss videre til de som var mest involvert i stordata prosjektene. Snøballutvalg blir brukt når vi har få ideer for hvordan man kan få tilgang til en gruppe (Oates, 2006). Gjennom disse kontaktene fra forstudien, fikk vi nye kontakter vi kunne spørre om å delta da vi gjennomførte intervjuene våren 2018. For SSB kontaktet vi statistikkavdelingen over e-post for å finne ut om noen var relevante for vår studie, som videresendte oss til den mest relevante informanten. Datatilsynet kontaktet via et åpent kontaktskjema, som videresendte oss.

Vårt kriterium for å velge informanter var basert på at de må ha innsikt og tanker om hvordan personvern vil bli håndtert i stordata prosjekter. Jo mer oversikt og involvering de har til personvern problemstillinger og stordata prosjekter, jo bedre.

3.3 Datainnsamling

Studien baserer seg på 14 intervjuer, der 3 intervjuer ble gjennomført høsten 2017, og 11 intervjuer ble gjennomført våren 2018. To av informantene i forstudien var også med som informanter i nye intervjuer i 2018; en fra NAV, og en fra Skatteetaten. Siden intervjuene var

utformet litt forskjellig, ble også svarene litt annerledes fra forstudien til masteroppgaven, men mange av spørsmålene dekket det samme.
Intervjuene var semi-strukturerte, der vi hadde en liste med tema vi ønsket å dekke, men åpnet opp for å endre rekkefølgen av spørsmål basert på hvordan samtalen gikk. Fordelen med semi-strukturerte intervjuer er at informantene kan snakke mer i detalj på de spørsmålene vi stiller, og introdusere utfordringer som de mener er relevante for temaet (Oates, 2006).

Vi inkluderte 39 faste spørsmål, der vi stilte oppfølgingsspørsmål til enkelte punkt der vi ønsket å lære mer eller for å få flere detaljer. I enkelte tilfeller ble spørsmål hoppet over, da svarene de ga var allerede besvart. Selv om ikke alle organisasjonene har kommet like langt med stordata, svarte de på spørsmål relatert til stordata problematikk. Med semi-strukturerte intervjuer får vi mulighet til å styre intervjuet ved at man er åpen for at samtalen kan ta uventede vendinger (Hellevik, 2003).

3.4 Transkribering av intervjuer

Samtalene ble transkribert i etterkant av intervjuene. Intervjuene ble tatt opp ved hjelp av mikrofon på mobil og laptop. Vi sørget for å spørre informantene i forkant om det var greit at vi tok opp intervjuene digitalt. I to av intervjuene ble vi bedt av informantene om å ikke ta opp samtalene, men heller notere ned hva personene sa. Grunnen til dette var at informanten ville være sikker på at informasjonen ikke kan benyttes til andre formål senere.

3.5 Litteratursøk

I forstudien høsten 2017 og gjennom skrivingen av oppgaven i 2018 har vi gjennomført litteratursøk med bruk av databasene Scopus, Oria og Web Of Science.
For vår litteraturgjennomgang definerte vi noen søkeord og søkekriterier. Vi søkte etter litteratur i form av artikler og bøker, og prioriterte artiklene etter hvor mye de var brukt av andre forskere ved å se hvor ofte de hadde blitt sitert. Vi fokuserte på artikler innenfor "informatikk" og "informasjonssystemer".

Noen av artiklene ble ansett for tekniske, og ikke så relevante som de andre artiklene i besvarelsen av forskningsspørsmålet. Vi utførte søkene ved å benytte oss av "OG" operatøren, for å finne de artiklene som var knyttet både til personvern, stordata og offentlig sektor.

Vi benyttet oss primært av søkeordene "stordata og personvern"; "offentlig sektor og personvern", samt "personvern".

3.6 Analyse og kategorisering

I første fase av vår analyse fulgte vi Oates (2006) sin metode for å dele opp dataene i tre ulike kategorier:

- Irrelevante segmenter som ikke var nødvendig for vårt forskningsspørsmål.
- Generell beskrivende informasjon relevant for forskningskonteksten.
- Relevante segmenter til vårt forskningsspørsmål.

Deretter benyttet oss av programvaren Nvivo 12 for å kode viktige poeng som kom frem i intervjuene. Dette er en prosess der man velger ut den teksten man ønsker å kategorisere ved å plassere dem i «noder». Vi kodet de sitatene som ble sett på som mest relevante, og plasserte informasjonen slik at de ble samlet under hver node. Informasjon som ikke var relevant til studien ble ikke kodet. Kategoriene våre var basert på eksisterende teorier i litteratur, og er kjent som en deduktiv tilnærming (Oates, 2006).

For å få en best mulig oversikt, kodet vi basert på stordata livssyklusmodellen, der vi plasserte relevant tekst under de kategoriene som passet innenfor innsamling, lagring og analyse. Informasjon som ikke passet under disse tre kategoriene, ble plassert for seg selv. Eksempel på dette var casebeskrivelse og informantenes arbeidsoppgaver.

Kategorier vi observerer i dataen, enten brukt av informanter eller i dokumentene, kalles en induktiv tilnærming. Dette vil tillate oss å ha et åpent sinn og la dataene "snakke" med oss (Oates, 2006). Etter å ha kategorisert ved bruk av en deduktiv metode ved å basere oss på livssyklusmodellen, kategoriserte vi videre i en induktiv metode, der vi spesifiserte underkategorier. De personvernprinsippene som ble tatt opp i tilknytning til GDPR ble sortert i sine egne kategorier.

Deretter søkte vi etter mellomforbindelser og temaer mellom kategoriene og segmentene. Dette gjorde det enkelt å se hvilke utfordringer som ble tatt opp flere ganger, og vi kunne dermed fastslå at disse var blant de mest gjennomtenkte og viktigste.

3.7 Validering av funn

Intern validitet handler om resultatene kan beskrives som korrekte (Yin, 2009). Ved å ta opp intervjuene sørget vi for at innholdet ble tatt vare på og transkribert, som hjelper med å validere funnene internt. Vi hadde mulighet til å dobbeltsjekke hvis det eksisterte en usikkerhet rundt hva som ble sagt. Ved å godkjenne forskningsdesign og intervjuguide sammen med veileder, bidro dette til å gi studien intern validitet (Hellevik, 2003).

Ekstern validitet handler om studiens resultater kan bli generalisert til områder utenfor case-studien som har blitt gjennomført (Yin, 2009). I utgangspunktet stilte vi de samme spørsmålene til alle informantene, som bidro til ekstern validitet. Det eneste unntaket var Datatilsynet, der vi formet en egen intervjuguide tilpasset dem. Grunnen til det er at de ikke er direkte involvert i stordata prosjekter på samme måte som NAV, Skatteeetaten og SSB. Ved å undersøke flere organisasjoner, kan dette potensielt bidra til ekstern validitet skulle man finne likheter og forskjeller mellom dem, og sammenligne med litteraturen.

3.8 Forskningsetiske retningslinjer

Vi søkte om mulighet til å behandle personopplysninger fra NSD - Norsk senter for forskningsdata. Vi fikk godkjent søknaden tidlig 2018, med krav om å slette all personidentifiserende informasjon etter at prosjektet var gjennomført. Vi hadde et ønske om å følge etiske retningslinjer, for å sikre respondentene rettigheter. Vi fulgte Oates (2006) sine retningslinjer for å sikre disse rettighetene:

Informanten kan velge å ikke delta
Skulle en informant ønske å ikke delta i intervjuet har personen mulighet til det. Deltagelse er informantens avgjørelse, og vi skal ikke presse noen til å delta om de skulle angre på å ha sagt ja.

Informanten kan trekke seg
Det er lov til å trekke seg fra intervjuet skulle informanten ønske dette. Det inkluderer også å kunne nekte å svare på visse spørsmål vi spør. Informantene har rett til å bli informert om intervjuets formål, før de gir samtykke til å delta i studien. Dette innebærer hva studien går ut på, hva vi ønsker å finne ut, og hva slags resultater som kommer til å skrives.

Informanten har rett til å bli anonymisert
Skulle respondenten ønske å bli totalt anonymisert er det også en mulighet. Vi har valgt å anonymisere informantene ved å ikke beskrive hvilke arbeidsoppgaver de har i prosjektene, eller ta med annen direkte identifiserende informasjon som navn og bosted.

Informanten skal gi et informert samtykke
Informanten har rett til å bli informert om studiens formål før intervjuet gjennomføres. Via digital korrespondanse har vi beskrevet:

- Hva studien går ut på, inkludert intervjuguide på forhånd hvis de ønsker dette.
- Hvem vi som forskere er og hvor vi studerer fra.
- Hvor langt intervjuet vil vare, og når studien fullføres.
- Hvordan data fra intervjuene blir behandlet.
- Om vi kan ta opp intervjuet eller om de ikke ønsker dette.

Rett til at data blir behandlet konfidensielt og trygt
Ved å benytte oss av NSD sine retningslinjer for konfidensiell behandling av personopplysninger sørger vi for å holde opplysningene om informanten trygt. Vi har benyttet oss av OneDrive som er en kryptert skylagringstjeneste utviklet av Microsoft, som vi begge får tilgang gjennom universitetet. På denne kontoen har vi lagret de transkriberte intervjuene og de originale lydfilene.

4. Forskningskontekst

4.1 Organisasjonene

4.1.1 Skatteetaten

Skatteetaten har som mål om å "sikre finansieringen av velferdssamfunnet". I etaten som er underlagt Finansdepartementet, jobber 6500 ansatte (Skatteetaten, 2018). Skatteetaten har ansvar for at skatter og avgifter blir innbetalt på riktig måte, i tillegg sørger de for at folkeregisteret holdes oppdatert. Deres visjon handler om et samfunn der "alle vil gjøre opp for seg". Det skal være enkelt å betale skatt (Skatteetaten, 2018).

For å nå hit kreves det stor tillit i befolkningen, blir det understreket av deres skattedirektør Hans Christian Holte. Viktigheten med tillit forsterkes med tanke på at alle innbyggere i Norges land vil komme i kontakt med Skatteetaten. Av den grunn er det viktig å opptre profesjonelt, og samtidig være nytenkende (Skatteetaten, 2018).

Deres strategi frem mot 2025 går ut på å møte fremtidens forventninger av samfunnet. De ønsker å bidra til en brukervennlig offentlig sektor. Skatteetaten skal se ut over sin egen etat, og bli enda bedre (Skatteetaten, 2018).

4.1.2 NAV

NAV, også kjent som Arbeids- og velferdsforvaltningen, ble etablert 1. juli 2006. NAV har så mange som 19 000 ansatte, hvor 14 000 av dem er ansatt i staten, og rundt 5000 av dem er ansatt i kommunene (NAV, 2018). Det er arbeids- og velferdsdirektoratet som har ansvar for å styre, lede og utvikle NAV. Direktoratet har ansvar for at NAV når de målene og resultatene som er satt, og i tillegg omsette politiske føringer til praktiske handlinger (NAV, 2018).

NAV sin oppgave er å forvalte en tredjedel av statsbudsjettet gjennom sine ordninger og ytelser, som blant annet er dagpenger, arbeidsavklaringspenger, sykepenger, pensjon, barnetrygd og kontantstøtte (NAV, 2018).

Hovedmålene til NAV inkluderer:

- Å få flere i arbeid, og færre på stønad.
- Et velfungerende arbeidsmarked.
- Å gi rett tjeneste og stønad til rett tid.
- Gi god service som er tilpasset brukerne sine behov.
- En helhetlig og effektiv arbeids- og velferdsforvaltning (NAV, 2018).

4.1.3 SSB - Statistisk sentralbyrå

SSB, kort for Statistisk sentralbyrå, er ansvarlig for å samle inn, produsere og publisere offisiell statistikk som relaterer til økonomi, befolkning og samfunn. Dette gjelder både på et nasjonalt, regionalt og lokalt nivå. SSB sin oppgave er å sørge for at innbyggerne i Norge kan ta beslutninger på grunnlag av statistikk som er pålitelig (SSB, 2018).

SSB henter data fra blant annet administrative registre og spørreundersøkelser. Det hentes også stadig mer informasjon direkte fra næringsliv og kommuner. De gjennomfører i tillegg intervjuer, enten via telefon eller ved å oppsøke folk i hjemmet (SSB, 2018).

SSB er underlagt Finansdepartementet, hvor styret er regjeringsoppnevnt. Gjennom statistikkloven fastslås det at SSB faglig sett er en uavhengig institusjon, men at de er underlagt de overordnede retningslinjene og finansielle rammene som regjering og Storting til enhver tid setter for virksomheten (SSB, 2018).

4.1.4 Datatilsynet

Datatilsynets oppgave er å føre kontroll med personvernregelverket. Datatilsynet har en tilsyns- og ombudsrolle. De gjennomfører kontroller hos virksomheter for å se om personvernreglene etterleves, og de jobber for at enkeltpersoner ikke skal bli krenket (Datatilsynet, 2018).

Datatilsynet ble opprettet 1. Januar 1980, og har omkring 50 ansatte som er fordelt på to fagavdelinger; administrasjonsavdelingen og kommunikasjonsavdelingen. De er et uavhengig forvaltningsorgan som er administrativt underordnet Kongen og Kommunal- og moderniseringsdepartementet (KMD) (Datatilsynet, 2018).

Datatilsynets hovedoppgaver er blant annet:

- Å kontrollere at lover for forskrifter som angår personopplysninger blir fulgt.
- Deltakelse i råd og utvalg.
- Å identifisere farer for personvernet.
- Gi råd og informasjon til publikum.
- Bidra til samfunnsdebatt om personvern (Datatilsynet, 2018).

4.2 Om prosjektene

4.2.1 Skatteetaten

Skatteetaten jobber med både rammeverk som Hadoop, og tradisjonell styringsinformasjon og analyse. Hadoop er åpen kildekode, og muliggjør kjøring av programvare og lagring av data på klynger av forskjellige datamaskiner (Johnsen, 2015). De jobber med maskinlæring og automatiserte algoritmer, og forteller at stordata blir brukt til å forbedre algoritmene, men at de er i et tidlig stadie på dette. Datavarehuset samler inn info til analyse for å finne strukturer og sammenhenger. De forteller at de er i starten for bruk av stordata.

En nøkkelperson for Skatteetaten sier at *"når det gjelder stordata er fokuset mest rundt modellering, metaanalyser og scoringsmodeller"*. De jobber nå med å utvikle prediktive modeller, det vil si modeller som kan forutse hvor de burde fokusere ressursene sine mest effektivt og kunne predikere hvem som gjør feil. Disse modellene brukes for eksempel til å beregne risiko som blir predikert basert på historikk fra konkrete hendelser, for eksempel ved levering av en selvangivelse, der de kan vurdere om det er feil på den. Målet er å sørge for å jobbe forebyggende slik at feil blir redusert til et minimum. Dette skal gjøres ved å bygge det de kaller for "etterlevelsespyramider", med etterlevelsesinndelinger med hvitt, grønt og rødt som brukes til å beregner risiko. De forteller at " *skal vi sørge for at de som ønsker å gjøre rett, men ikke har nok kunnskap eller informasjon til å gjøre det, får best mulig backing for å gjøre det. Også skal vi sørge for at de som ønsker å unndra skatt og klarer det blir stoppet som tidlig som mulig eller blir rettsforfulgt."*

Skatteetaten beskriver at hovedutfordringer med en risikobasert og kunnskapsbasert tilnærming er å bruke ressurser effektivt og håndtere det på riktig måte. De jobber med en ny plattform kalt Minerva som er i startfasen som erstatter dagens datavarehus når det kommer til personalstyring. De forteller at de har hatt ambisjoner om å bruke stordata i to-tre år på den nye plattformen, og at de nå jobber med å sikre at personvernet ivaretas. Dette skal være et prosjekt som skal foregå i 2018 og 2019, som de allerede har hatt et forprosjekt på. De antar at dette prosjektet vil vare til 2022/2023. De forteller at "*Steg 1 var jo å ha maskinvare og programvare som er i stand til å ivareta personvernet. Steg 2 er å bygge en fysisk plattform som støtter opp under personvernet. Steg 3 er å bruke de mulighetene når vi bygger opp tilgangsreglene"*.
Skatteetaten har flere tekniske prosjekter der de tilpasser systemene slik at de er klare for GDPR med tanke på tilgang og samtykke. Noen etater har kommet lengre enn andre, men de har mye klart til lovgivningen trer i kraft. Det er litt uklart å forstå alt fra juristenes side. De spesielle hjemlene fra Skatteetaten må tydeliggjøres for GDPR. I forhold til GDPR er det nye regler for innsyn, det er også krav om å gjøre minst mulig kopiering og duplisering (Datatilsynet, 2017).

4.2.2 NAV

NAV beskriver at med å koble ulike datakilder sammen og analysere data på tvers, åpner det seg en rekke muligheter for etaten for tjenesteutvikling, kunnskapsoppbygging og økt treffsikkerhet i dialog og oppfølging av brukerne.

Prosjektene beskrives av NAV slik:

- Nav ønsker å kunne sette inn ressurser mye tidligere til de som virkelig trenger det, og mindre ressurser for de som ikke gjør det. Ved hjelp av stordata og informasjon fra forskjellige kilder ønsker de å kunne predikere hvilke brukere som står i fare for å bli arbeidsledige. Dette skal være mye bedre enn de verktøyene og kunnskapen de har hatt tidligere, og sørge for å hjelpe allerede i starten av ledighetsløpet.

- De ønsker å forstå hvordan brukere opplever NAV. Når det gjelder tjenester som sykefravær, kan stordata gi informasjon om hvordan de kan begrense dette. Stordata kan bidra til å peke ut risikogrupper som NAV må rette sin innsats mot.

- Ved hjelp av stordata skal NAV øke sin kunnskap om dagens arbeidskraft, hvilken kompetanse brukere har i dag, og hva de kommer til å trenge i fremtiden. Gjennom stordata gis deg økt innsikt og forståelse for arbeidsgiveres situasjon, som skal bidra til å øke sannsynligheten for at brukere får jobb. De skal skape ny kunnskap om hvilke tjenester og arbeidsmarkedstiltak som fungerer for ulike grupper. Dette skal sørge for at NAV og brukere i større grad kan velge rett tiltak.

I tillegg ser de på mulighetene for å:

- Ta i bruk åpne data. De ønsker at forskere skal bidra med kunnskap ved å forske på NAV sine data. Dette kan for eksempel være en forskningsportal hvor man selv setter sammen data og kan gjøre statiske analyser.
- Lett forståelig visuell informasjon skal publiseres på nav.no.

4.2.3 SSB - Statistisk sentralbyrå

SSB forteller at de har ganske mye data, som sammenlagt går under begrepet stordata. I følge de er det først å fremst det å lage ukonvensjonelle data, eller bruke ukonvensjonelle data til statistikkformål, som utgjør stordata. SSB forteller at rent praktisk så er det tre saker som går inn under stordata. Den første er å kunne lage bedre og mer effektiv boligstatistikk, men også husholdningsstatistikk, hvor de ser på og måler strømbruk. De samarbeider med leverandøren for strøm som gir ut data. De forteller at de også jobber med et prosjekt innenfor prisstatistikk på data der de benytter seg av "web scraping". De ser på hjemmesider og fanger inn prisoppgaver. Et eksempel er en nettside med flypriser, der de samarbeider med firmaet som har satt opp siden. De forteller at flypriser er ustabile, og ser etter en klar nytte av å hente inn prisene umiddelbart når de endres. De kommer ikke til å kun basere seg på web scraping, da pris og statistikk er mer enn bare flypriser.

Det tredje prosjektet går ut på noe som kalles forbrukerundersøkelsen, der en husholdning lager en dagbok over inntekter og utgifter. I praksis blir dette veldig krevende, da man må skrive ned alt man kjøper i en viss tidsperiode. Dette blir også veldig kostbart, det blir stort frafall, og kvaliteten blir ikke nødvendigvis god nok. For to år siden begynte de med nye metoder, for å se på kvitteringsdata og transaksjonsdata fra kort. Utfordringen ligger i å kunne si noe om husholdningstyper og det mønsteret for forbruk de har, så de trenger å koble sammen kilder på ulike måter. Alle disse tre prosjektene er i startfasen. De har også utarbeidet et prosjekt som går ut på bruken av GPS for transportstatistikk. De har også kontaktet teleoperatører for å se på bruk av den type data, men de vet at det vil kreve store investeringer. De forteller at de er praktisk sett veldig tidlig i prosjektene.

4.3 Datakilder

Denne studien er utført hos NAV, Skatteetaten, SSB og Datatilsynet. Totalt ble det gjennomført 14 intervjuer med informanter på ulike nivå i organisasjonene. At både prosjektledere og ansatte i IT-avdelingene har deltatt gir et bedre bilde av hvilke personvernutfordringer en finner i offentlig sektor. Tabell x nedenfor viser oversikten av respondentene.

Tabell 3 - Oversikt over informantene

Organisasjon	Rolle	Arbeidsområde
NAV	Leder A	Tidligere prosjektleder
	Leder B	Nåværende prosjektleder
	Ansatt A	IT-avdelingen
	Ansatt B	IT-avdelingen
	Ansatt C	IT-avdelingen
	Ansatt D	IT-avdelingen
Skatteetaten	Leder A	Prosjektleder
	Leder B	Prosjektleder
	Leder C	Prosjektleder
	Ansatt A	IT-avdelingen
SSB	Ansatt A	Ansatt SSB
Datatilsynet	Ansatt A	Ansatt Datatilsynet

5. Resultater

I denne delen av oppgaven vil vi trekke frem funnene våre. Funnene vi presenterer er resultater fra 14 intervjuer gjort med organisasjonene. Våre funn belyses med sitater fra respondentene og dokumenter med informasjon om organisasjonenes stordata-prosjekter.

Inndeling er basert på stordata livssyklusen diskutert i litteraturkapittelet. Det vil si at respondentene har svart på personvernutfordringer innenfor stordata når det gjelder fasene innsamling, lagring og analyse. Innenfor hver fase trekker vi frem de mest sentrale funnene gjort i vår analyse av organisasjonene. Under intervjuprosessen ble det tatt utgangspunkt i en intervjuguide som var lik for alle organisasjoner.

Videre har vi trukket frem Datatilsynets anbefalinger og bekymringer når det gjelder hver fase. Datatilsynet fungerer som en kontrast til resultatene når det gjelder personvern. Siden de har en ombudsrolle når det gjelder bruken av stordata har det blitt tatt i bruk en separat intervjuguide.

5.1 Innsamling

5.1.1 Datakilder

NAV
Hos NAV er deres fagområder delt inn i siloer hvor hvert område er lovregulert når det gjelder hva de kan samle inn av informasjon. Det meste kan hentes gjennom data-avtaler og folkeregistre, ellers kreves det informasjon som brukeren må registrere selv i systemet. NAV henter inn informasjon om personer om den situasjonen de er i. Dersom man som arbeidsledig kommer inn til NAV vil data samles inn relatert til denne situasjonen. Dataene vil lagres i en database og gjøres tilgjengelig for veiledere. NAV samler også data fra bedrifter gjennom tjenesten "a-meldingen". Dette er data som inneholder opplysninger om inntekt og arbeidsforhold som må sendes inn av arbeidsgiver. Stordata er for NAV når man begynner å koble alle disse kildene sammen for å få innsikt man ikke visste var der. De ønsker i første omgang å ta i bruk data som allerede eksisterer i NAV-systemet, interne kilder, da de ser mye potensiale i dataene. Det er i første omgang strukturerte data, men NAV har også sett på å ta i bruk ustrukturerte metadata. Der er det likevel en vei å gå, understreker flere sentrale ansatte i organisasjonen.

NAV uttrykte at de sitter på enorme mengder med data. Leder A kommenterte at *"det er en utfordring for NAV for vi har så fryktelig mye data, så fryktelig mange datakilder. Hvis du kommer til oss og sier at jeg skal ha alle dataene dere har om meg, uavhengig av kilde, så har vi en liten utfordring der altså"*).

Skatteetaten
Skatteetaten samler inn mye data om enkeltpersoner, blant annet fra selvangivelsen. De har enorme mengder med data om enkeltpersoner, med strenge begrensninger på hvordan det brukes. Skatteetaten har hatt kontakt med Datatilsynet for å gjøre vurderinger for hva de kan bruke, etisk og strategisk ved innsamling av data. De har derfor en stor rettsavdeling som jobber med problemstillinger relatert til innsamling.

Skatteetaten samler inn data fra både interne og eksterne datakilder. Data som benyttes til stordata-plattformene er kopier av informasjonen de allerede har, dette inkluderer blant annet søknader, skattemeldinger, partsregister og annen strukturert informasjon.
Innenlands har Skatteetaten samarbeid med blant annet NAV og Tolletaten for å hente inn informasjon de trenger. De nevner tolldeklarasjoner som de bruker til analyse og finner mønstre fra. Leder A forteller at *"bankene rapporterer hvor mye individer har i banken, og hvor mye rente man betaler"*. Videre fører de også historikk på alle personer som kontakter dem.

Leder C sier at *"i all hovedsak vil nok Skatteetatens bruk av stordata fremover dreie seg om eksterne kilder"* (Leder C, Skatteetaten). Et eksempel på dette er Panama Papers, en enorm lekkasje med informasjon om verdensomspennende skatteunndragelse. Videre forteller Ansatt A at *"Vi får mye data fra andre land, de har til dels ganske dårlig kvalitet, mye er feilskrevet, vi forsøker å finne hvem det er"*.

Videre sier Skatteetaten at sammensetting av data kan være en utfordring. Leder B forteller at *"Hvis det blir endringer i innsamling av data, så man får brudd i tidsreiser, det er en stor utfordring vi har. Hvis ting endrer seg i tiden, det er en kamp. Brudd i tiden gjør det vanskelig å bygge modeller"*.

SSB:
Data hos SSB samles inn fra ulike undersøkelser de gjennomfører på bestilling av staten. I stordatasammenheng jobber de med prosjekter der de samler data eksternt via nettsider, kvitteringsdata og transaksjonsdata.

Når det gjelder bruk av stordata sier Ansatt A at *"informasjonsbehovet er større og kildene er ofte veldig endimensjonale. Informasjonsverdi fra et statistisk ståsted er begrenset. Du får data, men du får ikke hele sammenhengen. Det trenger vi jo også ganske mye av i analyse"*.

5.1.2 Samtykke

NAV
Når vi spurte NAV om samtykke, var de klare på at ettersom de var en offentlig organisasjon har de hjemmel til å samle inn data på flere områder. Dersom en bruker ønsker å ta i bruk NAVs tjenester er man pliktig til å oppgi opplysninger. I NAV-loven står det skrevet at opplysninger som samles inn om brukere kan tas i bruk til statistikk og analyse. Når det gjelder samtykker brukt til stordataformål var de fremdeles i en oppstartsfase, kunne sentrale nøkkelpersoner fortelle. Leder A sier at *"..dersom vi skal ta i bruk data som går utover vårt hjemmelsgrunnlag*

vil det være aktuelt å be om samtykker". Dette vil være tilfeller hvor data vil bli brukt i sammenhenger hvor en går utover primæroppgavene. En av informantene sa at det ville være sannsynlig at de kom i situasjoner hvor NAV ville være i tvil om hvor hjemmelsgrunnlaget lå. Det ville i en slik situasjon være aktuelt for NAV å be om et samtykke.

Skatteetaten
Med tanke på stordata og utfordringer knyttet til samtykke, forteller informantene at direktivet er en utfordring. I følge Leder C finnes det unntak for å bruke persondata; samtykke og hjemmelsgrunnlag. Videre kommenterer Leder C at *"..for Skatteetaten sin del så vil kanskje mer enn 99% av brukstilfellene våre komme fra hjemmelsgrunnlag. Hva Skatteetaten har lov til å samle inn og bruke av opplysninger, vil vurderes i hovedsak av hjemmelsgrunnlaget"*.

Ansatt A forteller om GDPR at *"nå blir det mye endringer med tanke på det nye direktivet. Vi har jo som etat spesielle hjemler, flere enn et ordinært selskap har. Vi får bruke data som er til skatteformål. Når det gjelder det juridiske så kan vi gjøre det. Spørsmålet blir hva er egentlig skatteformål"*.

Leder A sier at*"det er mange spesialregler for skattemyndighetene. Samfunnsoppdraget trenger ikke samtykke"*. Kun i spesielle tilfeller har de behov for direkte samtykke. Leder C gir et eksempel: *"Skal Skatteetaten gi tilbake informasjon om tidspunkt det ikke passer, så må vi be om telefonnummer. Og det er personopplysning, og da må vi bruke samme registreringsskjema, med en egen boks at vi må ha ditt samtykke til å lagre ditt telefonnummer"*. Slike data vil kun bli lagret i 24 timer før det blir slettet, da denne forespørselen blir behandlet innen denne tiden. Informanten forteller videre at *"med tanke på GDPR skal vi ha opplysningsplikt, ikke alle skal vite hva vi vet, det er vår måte å finne unndragelse på. Det er dilemmaer der, vi kan ikke være helt åpne med hva vi vet, hver kontroll skal finne hvem som gjør feil og hvem som er bevisst kriminelle"*.

SSB
SSB forteller at informasjonen samles inn frivillig, med et unntak som er arbeidskraftundersøkelsen som er obligatorisk. Denne undersøkelsen gir informasjon om hvor mange i befolkningen som er ute i arbeid. Her man pliktig til å svare hvis man blir trukket ut, hvis ikke kan en risikere å få bøter.

Datatilsynet
Når det gjelder samtykke så er Datatilsynet klare på at det er en utfordring. De ser det spesielt fra brukerens side, de har inntrykk av at brukere ikke ønsker å lese igjennom disse. *"Folk blir jo slitne av disse samtykkene. Man orker kanskje ikke lese igjennom, man trykker bare ja"* (Ansatt A, Datatilsynet). Videre informerer de om at det er viktig for organisasjoner å finne en balansegang mellom å gi nok informasjon, og samtidig ikke så mye informasjon at brukere ikke vil lese det. Datatilsynet råder organisasjoner til å ta i bruk illustrasjoner som bilder og ikoner for å gjøre samtykker mer forståelig og enklere for folk å ta stilling til. Det er også et problem at samtykkene ofte er vanskelige å forstå. *"Spesielt dersom de er skrevet i et juridisk språk"*, legger Ansatt A til. Videre kan de forklare at dette blir skjerpet inn med den nye lovgivningen, GDPR.

Datatilsynet anbefaler organisasjoner å ta i bruk deres veiledning for hvordan samtykker bør se ut.

Datatilsynet ser spesielt 'take it or leave it'-samtykker som utfordrende. Det vil si at organisasjoner krever at flere vilkår skal være godkjent før brukere får lov til å ta i bruk tjenesten. Datatilsynet forklarer videre at med det nye personverndirektivet som innføres så vil dette ikke lengre bli tillatt. *"Med den nye loven så må folk i større grad få lov til å velge selv. Du kan få lov til å bruke denne appen her, og hvis du vil så kan du avlevere informasjon om din lokasjonsdata. Eller du kan velge å la være. Man skal selv kunne velge hvilken informasjon man ønsker å dele, og ikke"* (Ansatt 1, Datatilsynet).

5.1.3 Dataminimering

NAV
Knyttet til innsamling og personvernet ligger det utfordringer for NAV når det gjelder prinsipper om dataminimering. De gav inntrykk av at det legges press på organisasjonen rundt disse prinsippene. NAV skal kun hente inn data de trenger for å kunne gjøre akkurat det som brukeren har krav på. En informant kunne meddele at dette bryter med deres ønsker om å lage gode tjenester for brukerne deres. Ansatt C sier at *"vi kunne lagd mye bedre tjenester for brukerne våre dersom vi kunne hatt et større datasett"*. I tillegg kunne en de meddele at det var vanskelig å gjøre vurderinger på om de hadde lov til å ta i bruk dataene i utgangspunktet. I et stordataperspektiv gjøres det vurderinger på om de har behov for dataene, eller om de klarer seg uten. Ledelsen kunne blant annet fortelle at dataminimering ikke var et fokus før GDPR kom på banen. Da vi spurte dem om hvordan de lå an i forhold til det nye personverndirektivet kunne de fortelle at det var i en oppstartsfase. De har satt seg inn i hva lovverket har å si for stordatatjenestene og har vært påpasselig med at det de gjør er i tråd med loven. De har blant annet kjørt et stordataprosjekt som er helt i tråd med det nye direktivet. *"Vi ønsker tillit fra brukerne, og vil ikke presse grenser"* (Leder A, NAV). Samtidig har de hatt flere runder med Datatilsynet for å passe på at prosjekter har blitt utført i tråd med lovverket.

Skatteetaten:
For Skatteetaten blir all data som er samlet inn brukt til noe, men data blir ikke spesifikk samlet inn for stordatabruk.

Leder A forteller at *"All vår datainnsamling bruker vi alltid til noe. Vi har ikke noen data som ligger rundt og aldri blir brukt"*. De nevner likevel at det ikke nødvendigvis blir skapt handling ut i fra all informasjon. *"Vi får jo inn rundt 400 000 mva oppgaver. Så vi ser jo ikke på alle, de vurderes jo maskinelt. Om det er korrekt informasjon, om det er en risikoscore. På den måten bruker vi jo informasjon. Men vi gjør kanskje ikke direkte noen action ut av all informasjonen"* (Leder A, Skatteetaten).

De forteller også at med hensyn til personvernforordningen (GDPR) har satt et nytt fokus på hva de har hjemmelsgrunnlag til å benytte til videre bruk. Før personvernforordningen ble et fokus

for Skatteetaten tok de i bruk andre begrepet. *"En gang ble det sagt, det var før personvernordningen kom, så brukte man begrepet alt vi vet er alt vi gjør. Det begrepet bruker vi ikke lenger. Fordi vi ikke har lov til å bruke alt vi vet til alt vi gjør. I tillegg så var det nok vi som lanserte begrepet kanskje på et litt overordnet synspunkt og nivå, og tenkte ikke over de ganske strenge begrensningene vi har. Nå blir plutselig det med at man har lov til å bruke data et tema i alle andre virksomheter også. Så vi er velkjent med den problemstillingen"* .

Når vi spør han om de samler inn data for senere bruk svarer de nei. Det er strenge regler dersom Skatteetaten skal samle inn data. Videre forklarer de at de ikke kan samle inn data bare fordi de syns at det kan være interessant. Selv uttrykker de at dette er den store ulempen ved at være en statlig etat, men at det samtidig er viktig for forbrukeren. Hver gang de ser en potensiell datakilde tar de i bruk jurister. De må ha hjemmelsgrunnlag for å ta i bruk opplysningene.

SSB:
SSB samler kun inn data hvis de har et statistikkformål for det. De forteller at det kan gjøres på mange ulike måter i dagens lovverk. Statistikkloven pålegger SSB at de ikke skal samle inn noe data som de ikke behøver. Ansatt A sier *"vi har det her som vi kaller oppgavebyrde, og fremfor alt opp mot foretak, men også mot individer. At vi skal ikke stille unødvendige spørsmål*. De sier at de ikke samler inn data som de kan trenge frem i tid.

SSB følger to prinsipper;
1) de må ha et formål med datainnsamlingen,
2) formålene må begrense datafangsten til så mye som mulig.

Informanten forteller at *"vi må optimere balanse mellom formål og datainnsamling"*. Ved innsamling av stordata vil det koste mye, så de trenger å være sikre på at de faktisk trenger dataene de samler inn.
Et eksempel de gir er hvis de skal samle inn data fra en aktør. Ved å kun hente inn de spesifikke dataene, om for eksempel informatikkstudenter i Norge, minsker SSB risikoen for at data kan komme på avveie og at noen kan lide skade.

Datatilsynet
Datatilsynet ser på prinsippene om dataminimering som utfordrende når det gjelder stordata. Ansatt A sier at *"..da stordata handler om å samle inn så mye data som mulig for å få innsikt, blir det en motsetning til prinsippene våre om dataminimering"*. Applikasjoner som for eksempel skal kunne hjelpe deg med reiseplanlegging skal ikke hente inn data fra kontaktlister og lignende.

Når det gjelder stordata og dataminimering synes Datatilsynet at det er problematisk at organisasjoner kan sitte på det de omtaler som ekstreme mengder med data. Spesielt med tanke på at organisasjoner kan bygge profiler på enkeltindivider. De mener at dette kan skape en maktubalanse i samfunnet. *"Det har personvernutfordringer fordi man kan få maktubalanse mellom enkeltindivider på den ene siden, og store virksomheter på den andre. Som sitter med en enorm detaljkunnskap som gjør at de kan få et overtak over enkeltindivider"* (Ansatt A, Datatilsynet). De formidler at det er problematisk at enkeltindivider ikke vet hvor store

informasjonsmengder som organisasjoner kan ha lagret. Videre peker de på den nye lovgivningen når det gjelder dataminimering, som sier at den enkelte skal ha større kontroll over sine egne data. En skal som bruker ha innsikt i hvordan disse dataene behandles.

5.1.4 Formålsbegrensning

NAV
NAV gav inntrykk av at de hadde klare hensikter bak bruken av stordata. De samler ikke inn data for stordatas skyld, det ligger et formål bak bruken av det. De samler heller ikke inn data som kan bli brukt til senere bruk. All data som er samlet inn og som ligger lagret hos NAV er der for en grunn.

Dataene samles inn for å yte de tjenestene som NAV skal gjøre for brukerne sine. De samles inn fordi de skal brukes i forbindelse med enten å ta stilling til om en bruker har krav på en ytelse, eller til å beregne en ytelse. Leder A kommenterer at *"vi driver ikke med data for datainnsamlingens skyld, for at vi kanskje skal bruke det til noe annet senere. Det ville man kanskje ha gjort hvis forretningsideen var å finne opp helt nye tjenester basert på stordata, men det er ikke vår kjerneoppgave"*. De kan imidlertid legge til at de tar i bruk sosiale medier til kontrollformål.

Når det gjelder å ta i bruk data fra eksterne kilder som sosiale medier så var dette noe som ikke var aktuelt for NAV per tidspunkt. En nøkkelpersonene i stordata-prosjektene kunne legge til at de ikke ønsket å være frempå når det gjelder dette området. De ser heller potensialet i data som allerede ligger i NAV-systemene. Leder B sier at *"NAV skal ikke være helt i front på det området, vi er ikke den som tester grensene for mye. Jeg tror ikke vi kommer dit med det første, vi har så mye data internt som vi ikke har brukt hele potensialet. Så får heller andre jobbe med sosiale medier"*.

Videre kunne ansatte i NAV fortelle at formålsbestemthet kunne være utfordrende. De som jobber med den tekniske tilnærmingen av stordata har måttet spørre seg selv om hvilke data de har lagret, og hvordan de kan bruke dem. I henhold til den nye personvernloven ligger det en del uavklarte elementer for NAV.

Skatteetaten
Skatteetaten sier at det ligger en grad av formålsbegrensning i hva de kan bruke data til. De kan bruke data relatert til skatteformål. Enkelte typer data kan de ikke bruke til modellering for stordata. Leder C sier at *"spørsmålet blir hva er egentlig skatteformål. Vi har jo statens innkrevingssentral som krever inn penger. De er en del av Skatteetaten. Dette gjelder for eksempel data om bøter og NRK-lisenser. Innkreving, data om hva folk skylder, får vi ikke bruke"*. Skatteetaten har et samtykke så lenge dataene de bruker gjelder skatt.

Ansatt A forteller at lover og regler begrenser bruken av stordata; *"lover begrenser i aller høyeste grad. De begrenser hva vi har ønsker om å bruke og behov for å bruke. Hvis du tegner et*

mengde diagram, kan du si at det vi har behov for utgjør 100% av mengden, så er det vi har hjemmelsgrunnlag for å bruke kanskje 20%".

De forteller at når det kommer til å gjennomføre prosjekter må det være både hjemmelsgrunnlag og kapasitet i etaten. Leder C sier at *"det første steget er at man melder ifra om det er et prosjekt. Når prosjektet er over vil det være en på forretningssiden som tar forespørselen videre. Til slutt vil det bli foretatt en juridisk vurdering om det finnes juridiske unntak for å ta i bruk disse opplysningene"*.

Det kan være et trangt nåløye å gå igjennom. Det må være juridisk grunnlag for å få lov til å ta i bruk dataen, noe som kan medføre en lang juridisk prosess. I enkelte tilfeller innebærer dette å bidra til at myndighetene endrer loven. *"Hva skal til for at vi får hjemmelsgrunnlag for disse opplysningene, slik at vi kan utføre samfunnsoppdraget vårt bedre? Da vil det ofte innebære en endring i lovverket*" (Leder C, Skatteetaten).

Etiske vurderinger kan gå over hva som er lovlig, det ikke er slik at alt som i utgangspunktet er lovlig å gjøre, også er det etiske riktige. Leder A forteller at *"en problemstilling jeg blir minnet på er at vi ikke bare ta stilling til hva du har juridisk lov til å bruke. Du skal også ta stilling til hva er etisk forsvarlig å gjøre. I mange tilfeller så vil Skatteetaten ha juridisk grunnlag for å ta i bruk et sett informasjon. Men om vi gjør dette bryter vi med den jevne borgers følelser, altså dette burde vi ikke gjort, eller her går vi for langt"*.

SSB

Rutinene for å sikre personvern i stordata er lik deres praksis når det gjelder andre type data. Det baserer seg på det informanten beskriver som "need to know basis", man skal ikke vite mer enn det man trenger for å utføre oppgaven sin.

Informanten forteller oss at *"dataen får ikke bli brukt til noe annet enn det formålet som er fastslått fra begynnelsen. Ser man en annen nytte av en undersøkelse som blir gjort for ti år siden. Da får man ikke bruke dem, for det var ikke samtykke da. Man må respektere de beslutningene som ble tatt"*. Informanten sier at det kan være uheldig at det ikke finnes fremtidig nytte for disse dataene.

Måten dataene håndteres på er via rutinene de har. En utfordring som beskrives er hvis brukere motsier seg fra å gi fra seg data, og hvem som kan avgjøre hvilke data kan brukes i andre sammenhenger.

Datatilsynet

Når det gjelder prinsippene om formålsbestemthet er det viktig at det er et formål bak innsamlingen og bruken av dataen. Datatilsynet tar frem et eksempel for å illustrere dette. *"Dersom en skal lage en applikasjon og den har til formål å vise deg hvor den nærmeste busstasjonen er, så kan de samle inn lokasjonsdata. Det har de et formål med. Det blir derimot utfordrende dersom en begynner å gjenbruke data til nye formål"*(Ansatt A, Datatilsynet). Datatilsynet forteller at det er utfordrende når det gjelder loven, da en som borger ofte ikke samtykker til denne nye bruken av dataen.

Ansatt A kommenterer at *"om vi er på Twitter så har jo vi samtykket til hvordan denne dataen blir samlet inn. Men hvis for eksempel et forsikringsselskap går inn og bruker disse opplysningene som jeg har lagt igjen på Twitter i en analyse av meg, så strider dette mot dette formålsprinsippet. For jeg har aldri samtykket til at forsikringsselskapet skal kunne bruke mine data til det formålet. Det er videre bruk av dataene som ikke er tillatt"*.

Videre forklarer de at en alltid må ha et lovlig behandlingsgrunnlag for å samle inn opplysninger. Det vil si at man ikke bare kan ta opplysninger som er tilgjengelige og gjenbruke de uten å ha et samtykke.

Tabell 4 - Utfordringer knyttet til innsamling av stordata

Organisasjon	Datakilder	Samtykke	Dataminimering	Formålsbestemthet
NAV	Interne strukturerte datakilder. Kilder som i stor grad blir registrert av brukere selv. Ustrukturerte data vil bli aktuelt senere. Største utfordringen er mengden av data.	Samtykke trengs ikke så lenge de har hjemmelsgrunnlag. Samtykker kan bli aktuelt i fremtiden	Skal kun hente inn data de trenge for å gjøre det brukeren har krav på. Kunne lagd bedre tjenester dersom de hadde hatt et større datasett.	Skal være en hensikt bak bruken. Kan være utfordrende. Til tider vanskelig å fastslå hva de kan bruke til stordataformål.
Skatteetaten	Interne og eksterne datakilder. Data som samles inn og som registreres av brukere selv. Største utfordringen er brudd i tidsreiser.	Trenger ikke samtykke dersom det er i forbindelse med samfunnsoppdrag. Samtykke trengs i spesielle behov.	Henter kun inn data de har bruk for. Ser på det som en ulempe for etaten, men en fordel for brukerne.	Kan bruke data relatert til skatteformål. Det de kan bruke utgjør bare 20% av behovet de har. Må ligge et juridisk grunnlag der. Gjør også etiske vurderinger.
SSB	Interne og eksterne datakilder. Data samles inn på bestilling av staten. Største utfordringen er at kildene er endimensjonale.	Informasjon samles inn frivillig. Deltar man, samtykkes det.	Samler kun inn data dersom de har statistikkformål for det. Samtidig må de begrense datafangsten så mye som mulig.	Lik praksis som med konvensjonelle data. «Need to know»-basis. Data får ikke benyttes til noe annet enn det formålet som er fastslått fra begynnelsen.

5.2 Lagring

5.2.1 Hvordan data lagres

NAV
Når vi begynte å snakke om lagring ble vi fort gjort oppmerksomme på at informantene var tilbakeholdne med å dele opplysninger. Spesielt ansatte ved de tekniske avdelingene uttrykte bekymring når vi spurte om emnet. Det vi fikk vite var at lagring av data foregår i Norge, hvor det blir tatt i bruk datavarehus. De kunne videre forklare at de har strenge rutiner når det gjelder tilgangskontroller knyttet opp mot lagring. For å få tilgang til systemene til NAV fikk vi blant annet vite at en må være norsk statsborger, selv hvis en jobber som ekstern konsulent.

Skatteetaten:
Lagring hos Skatteetaten gjøres i et lokalt datavarehus. Leder B forteller at *"Vi har alt inhouse. De aller fleste skatte-organisasjoner har jo egne systemer. Alt dette er privat for å kunne ivareta sikkerhet"*.

Ansatt A forteller oss at det finnes en del restriksjoner på hva som kan lagres og hvor lenge. *"Vi må ta vare på dokumentasjonen i tilfelle noe må tas opp i retten. Det har vært tilfeller hvor vi har tatt vare på det i mer enn ti år. Med ny skattelov så er fem år den vanlige tiden, men i grove tilfeller kan det være ti år. Det blir også lagret så lenge det er relevant. Hvis det ikke kan ankes mer og konklusjon er endelig, så rydder vi dataene. Samtidig er det nesten aldri mulig å slette 100%. Det som ligger på backup er i praksis nesten umulig å fjerne"* (Ansatt A, Skatteetaten).

Datatilsynet
Når det kommer til lagring er det viktig at sensitive opplysninger lagres kryptert, sier Datatilsynet. Samtidig er det viktig at en har et formål med å lagre opplysningene. Videre kommenterer de at dersom formålet er oppfylt, må dataene slettes. Datatilsynet går også på tilsyn for å sørge for at dette følges opp av organisasjoner. I stordata-sammenheng er det mange som ønsker at data skal lagres så lenge som mulig, for da kan de lære mest om kundene sine, forteller de. *"Da må vi ofte gå inn å si at det blir for lenge"* (Ansatt A, Datatilsynet).

5.2.2 Skylagring

NAV
For NAV sin del er det ikke per tidspunkt aktuelt å ta i bruk skylagring. Når det gjelder stordata-prosjektene kom det frem at det var flere elementer de ønsker å få på plass før de kan tenke på å lagre data i skyen. Da vi spurte informantene hos NAV om hva de syns om skylagring og personvern var det klart at det var ulike meninger om dette. Ansatte som jobbet med tekniske aspekter av stordata-prosjektene så på det som et element med høy risiko, mens andre hadde meninger om at dataene i skyen ville være enda sikrere. Samtidig kom det frem at NAV ønsker å

ha kontroll over hvem som har tilgang til dataene. Når det gjelder fremtiden ville de ikke utelukke bruken av det, da de var klare på at skylagring var en viktig del av stordatas fremtid. En av informantene vi intervjuet kunne blant annet meddele at han skulle delta i et møte om akkurat denne problemstillingen.

Skatteetaten
Skatteetaten mener det ikke er aktuelt med skylagring på grunn av sikkerhetsutfordringer. Maskinvare og programvare og den fysiske plattformen må være i stand til å ivareta personvern for prosjektet. Leder C forteller at *"Det er ikke aktuelt for Skatteetaten for noen systemer som inneholder person eller organisasjonsdata. Vi bruker det litt i forbindelse med 'proof of concept'. Men da er det kun anonymiserte eller på annet vis ufarlige data"*.

SSB
Skylagring er ikke aktuelt for SSB nå, men de sier at det i fremtiden kan være aktuelt med ulike skylagringsløsninger.

Datatilsynet
Dersom en tar i bruk tredjeparter når en skal lagre data i skyen er det viktig at virksomheten er sikkert på at dataene er trygt lagret det, forteller Datatilsynet. Ansatt A kommenterer *"Hvis kommunens eller barnevernets opplysninger ligger i den skyen så må jo kommunen være sikker på at den skyleverandøren kan garantere sikkerhet"*. For at man skal være sikker, er det ønskelig fra Datatilsynet at tredjeparten er innenfor europeisk juridisk seksjon. De kan blant annet fortelle at det er leverandører som har egne datasentre i europa for å betrygge organisasjoner.

5.2.3 Tilgangskontroll

NAV
NAV kunne blant annet meddele at de har robuste systemer for hvem som skal ha tilgang til hva. Dataene i NAV sorteres etter fagsystem, hvor det er dataeiere i hvert system som har et overordnet ansvar over dataene. Rundt hvert system er det egne team hvor hvert system har sine egne databaser. Her sitter det også utviklere. *"Dersom man som ansatt i NAV er utvikler innenfor systemet som håndterer dagpenger har en kun tilgang til den databasen og de dataene som eksisterer der"* (Ansatt A, NAV).

Dersom et team er avhengig av informasjon fra et annet team, samarbeides det. Det er ikke tillatt å dykke ned i hverandres data. Innenfor IT-avdelingen jobber det også 'Data-scientists'. Dette er ansatte som kan hente ut data fra flere ulike kilder, og som har vide tilganger. De kan hente opp data fra ulike kilder og løfte det opp på et mellomlag. Det er imidlertid veldig få i organisasjonen som kan se alt, det er stort sett uidentifiserbar informasjon. De som har spesialtilganger i NAV har tilgang til både pseudoanonymiserte data og data med fødselsnummer. Tilgang til data er styrt etter hvilke roller man har, hvor en vanlig bruker har kun tilgang til anonymiserte og aggregerte data.

NAV har også logging i sine systemer. Systemene har logging av alle arbeidsrutiner som utføres. Det er mulig å gå tilbake å se på hvem som har sett på hvilke data. En informant kunne fortelle om tilfeller hvor det har vært saker hvor brukere har bedt om innsyn i hvem som har tilgang til hvilke data.

Skatteetaten

Leder C forteller at med 7000 ansatte kan det være et problem med for bred tilgang til dataen med tanke på personvern og sikkerhet. *"Det er veldig få som sitter og jobber med modellering. Hvis man ser på personvern som sikkerhet og tilgang til dataen så har det vært et problem med at det har vært litt for brede tilganger"*. Informanten forteller videre at *"Det er vi veldig nøye på er at det er akkurat de som trenger dataen, innenfor akkurat det tidspunktet, som får tilgang. Vi har ikke noen som kobler bare data for at det ser gøy ut"* (Leder C, Skatteetaten).

Innenfor den nye plattformen Minerva vil det være krav om bruker-ID og informasjon om hva de skal gjøre beslutter tilgangen de har innenfor plattformen.

Skatteetaten nevner flere dimensjoner på tilgangsstyring:
1. Hva man skal gjøre og hva man får lov til å lage av analyser, om man kan bruke det andre har.
2. Hvilket emneområde man skal inn på, som økonomi, data, og lignende.
3. Hvilken sensitivitetsnivå man skal ha tilgang til.
4. Organisatorisk tilknytning.

Tilgang styres ut i fra hvor i organisasjonen man jobber, dette skjer automatisk. Disse har vært i bruk siden 2014. Tilgang kan styres ved at du sier hva du skal jobbe med, hvilken rolle du skal ha og lignende. Leder B forteller at *"..og da er selvfølgelig fremdeles muligheten for at denne brukeren kan jukse, og benytte seg av en rolle som er feil. Men det er den typen integritet vi faktisk forventer av ansatte"*. Som ansatt hos Skatteetaten skal man ikke ha flere eller bredere tilganger enn det som er nødvendig for å utføre oppgaven. Et viktig element er at informasjon om bruken av systemet logges.

SSB:
SSB benytter seg av en "need to know" metode i tillegg til å ha en egen datafangst-avdeling. Det innebærer at man ikke har mer tilgang til informasjon enn det som er nødvendig for å gjennomføre oppgaven. Det finnes alltid en eier til dataen som står med ansvaret, forteller de. For å få tilgang til data må man søke om å få det, dette er veldig styrt. De benytter seg av tydelige rutiner, som også vil gjelde for stordata.

Datatilsynet
Når det gjelder tilgangsstyring har Datatilsynet regler en må forholde seg til, spesielt ved bruk av sensitive opplysninger. Datatilsynet har hatt flere saker der folk har vært inne å sett på informasjon om kjendiser. Når det gjelder tilgangsstyring er det viktig at det bare er det som har et tjenestebehov som skal ha tilgang til dataene. Videre får de frem viktigheten av loggføring, at alle oppslag bør loggføres. Datatilsynet sier at dette er en spesiell ting som de er ute å ser etter når de er på tilsyn innenfor offentlig sektor.

“Det skal jo ikke være sånn at man kan snoke i naboens NAV profil. Det er jo kjempeviktig i det offentlige å ha gode systemer for det” (Ansatt A, Datatilsynet).

5.2.4 Sensitivitet

NAV
For NAV så er sensitive data blant annet data som kan si noe om en persons religion og legning, det som er definert i lov. De er også opptatt av personidentifiserbar data. Ansatte kan ha behov for å jobbe med sensitive data, men NAV er ikke bekymret over disse dataene da alle ansatte har signert erklæringer. *“Når det gjelder å bruke denne dataen til analyseformål kan det være problematisk”*, forteller Leder A. NAV har et eget personvernombud som jobber med disse utfordringene sammen med deres juridiske avdeling.

Skatteetaten
Skatteetaten forteller at “alt” er vurdert som sensitive data, men det er ulik grad av sensitivitet. Alt som kan knyttes til en person er i utgangspunktet sensitive data. Informasjon fra folkeregisteret er et annet eksempel på sensitiv informasjon.

Skatteetaten jobber med ulike grad av sensitivitet på data som lagres. Dette er kodet i tall for å beskrive hvilket nivå av sensitivitet det er snakk om.Kode 6 og 7 er for eksempel adresse, der kode 6 beskriver de som har beskyttet adresse. Geolokalisering og informasjon om barn er også beskyttet informasjon. Ansatt A sier at*“Det som ikke kan ikke bli eksponert er adresse, geologaklisering, barn. Dette er høyt på radaren for det kan stå på helse”*.

Også annen type informasjon kan være sensitiv, selv om det ikke dreier seg om personvern, som for eksempel regnskapsinfo og kontrakter som er sensitivt i forhold til aksjehandel.

SSB
SSB ser på sensitive data som data som kan bli brukt til å identifisere personer. SSB samler inn data på et overordnet nivå. Ved spørreundersøkelser samtykker brukeren selv til datainnsamlingen ved å gjennomføre den.

Tabell 5 - Utfordringer knyttet til lagring av stordata

Organisasjon	Lagring	Skylagring	Tilgangskontroll	Sensitivitet
NAV	Foregår lokalt i Norge. Tar i bruk datavarehus.	Skylagring ikke aktuelt. Delte meninger om skylagring er en risiko. Ser ikke bort ifra fremtidig bruk.	Kun tilgang der man har et tjenestebehov. Tilgang styrt etter rolle. Dersom ansatte behøver data fra en ekstern avdeling samarbeides det. Få som har tilgang til alt. Tar i bruk logging.	Sensitive data er det som er definert i lov; religion, legning o.l. Har personvernombud som takler utfordringer med sensitive data.
Skatteetaten	Foregår lokalt i Norge. Tar i bruk datavarehus. Strenge rutinger for hvor lenge data kan lagres.	Ser på skylagring som uaktuelt på grunn av sikkerhetsutfordringer.	Automatisk tilgangskontroll etter rolle. En risiko at det kan bli gitt for brede tilganger. Forventer integritet av ansatte.	Vurderer "alt" som sensitive data, men det er ulik grad av sensitivitet. Informasjon om barn og personer som lever på beskyttet adresse er mest sensitivt.
SSB	Tradisjonell lagring.	Skylagring ikke aktuelt nå, men kan være det i fremtiden.	Tilgang etter «need to know»-basis. Ansatte skal ikke ha mer tilgang enn det som er nødvendig for å gjennomføre oppgaven.	Sensitive data er data som kan bli brukt til å identifisere personer.

5.3 Analyse

5.3.1 Anonymisering

NAV
De som jobber med analyser har ikke tilgang til for eksempel fødselsnummer og personnummer. Det er kun to ansatte i hele etaten som kan gjøre koblinger med slik data. Når det gjelder anonymisering så har det byttet ut alle fødselsnummer og personnummer med en egen NAV-ident. Det er et nummer som ikke betyr noe. Alle personer som er registrert i NAV-systemet har en NAV-indent. Identen kan ikke spores tilbake til enkeltpersoner, men den er lik gjennom hele NAV-systemet. Dersom man har både arbeidsledighetstrygd og barnetrygd kan indenten koble det, men man vet ikke hvem som står bak indenten. De kunne si at dersom man driller hardt nok ned i dataene så vil man til slutt stå igjen med en person. De vil aldri gå ut med denne informasjonen offentlig. Dette gjelder både internt og eksternt. Hvis de skal publisere noen tabeller, så blir alt som har mindre fem i en celle maskert.

Når vi spurte NAV om tiltak som ville være aktuelle å iverksette med tanke på anonymisering kunne de fortelle at de vurderer å ta i bruk støyfilter. Leder A sier at *"Dette vil si at alle individdata vil pålegges litt støy, hvis du er 26 år gammel vil det i databasen stå at du er 25,7"*. Fordelen ved bruken av dette er at summen av dataene blir riktig, men at hver for seg fremstår dataene som ufullstendige.

Skatteetaten
En utfordring er å anonymisere data tilstrekkelig. Den som jobber med dataen skal ikke kunne vite hvem det er dataen handler om. En som analyserer stordata går bredere enn en enkelt saksbehandler. Når det kommer til utvikling av prediktive modeller sier Skatteetaten at de ikke trenger å vite hvem individene er, derfor er anonymisering sentralt. Leder A forteller at *"Da er anonymisering et tiltak i forhold til bygging av modeller, du kan man ta bort problemet ved at det ikke er identifiserbar info du jobber på. Du kan jobbe på data som er anonymisert, den som jobber kan ikke se hvilken person det er"*.

En utfordring er at modeller har gjort systematiske feil. Prosessene for å teste modeller er ganske gode, så ved store feil har det blitt oppdaget. Skatteetaten jobber med "privacy by design" for den nye plattformen, Minerva. Det innebærer at leverandører må levere en løsning som støtter opp under dette.

En utfordring er i det å kunne automatisk søke gjennom tekstdokumenter for å fastslå om det finnes identifiserbare personer i dokumentet eller ikke. Kombinasjoner av opplysninger kan være identifiserende. Med for mye granularitet kan det bli vanskeligere å få noe ut av det, spesielt video, bilder og lydfiler.

"Å kunne enkelt koble en sammenheng med andre opplysninger i datasettet fra 35 til 50 ulike kilder, er en utfordring", sier Leder C. Videre forteller informanten at en spesiell utfordring

knyttet til stordata og personvern er at verktøy for å bevise at det finnes identifiserbare personer i opplysningene kun vil være villedende. *"Utfordringen ligger mye på at jo mindre man kjenner innholdet, altså jo mindre metadata man har om denne såkalte kilden, jo vanskeligere er det å sette gode tilgangsbegrensninger på det. Da kan det gå utover personvernet. For eksempel i den grad vi har dokumenter liggende, en del kan løses med verktøy som kan søke igjennom og identifisere, ,en det vil i beste fall være veiledende. Det vil ikke være helt nøyaktig"* (Leder C, Skatteetaten).

SSB
Ansatt A beskriver en prosess der *"..så fort de har gjort de koblingene de trenger på grunn av statistiske behov, så fjernes identifiseringen ganske raskt"*. Slik unngår de at informasjon blir spredt.

Når det gjelder personvern knyttet til enkeltpersoner, jobber SSB først og fremst med å skape statistikk på et overordnet nivå. De er ikke spesielt opptatt av enkeltpersoner, men av større grupper. De forteller at de jobber fra enkeltindividene og opp i det store, mens for eksempel Skatteetaten i kontrast jobber ovenfra og ned på enkeltindivid nivå. Dette gjør at de er litt annerledes enn andre offentlige organisasjoner.

Datatilsynet
Når det gjelder faren for at brukere kan bli identifisert så anbefaler Datatilsynet at organisasjoner tar i bruk gode teknikker for anonymisering. At en tar i bruk de mest solide og robuste metodene. Datatilsynet opplever misforståelser når det gjelder anonymisering av data. At man tror at de er anonyme, men i virkeligheten har man bare pseudo-anonymisert dem. En har tatt bort navn, adresse og telefonnummer, men som ligger igjen av data er ofte nok til å identifisere individer. *"Hvis man har yrke, kommune, kjønn og alder, da skal det ikke så mye til da før man klarer å re-identifisere noen. Min sjef pleier å si at det skal bare to datapunkter til for å identifisere han. Det er at han heier på et bestemt fotballag, og at han eier en katt som er av en veldig spesiell rase"* (Ansatt A, Datatilsynet).

5.3.2 Datakvalitet

NAV
For personvernutfordringer knyttet til analyse så er det viktig å sørge for at personidentifiserende data ikke kommer ut. Det er viktig å sørge for at de resultatene man får av analysene på stordata er korrekte. NAV kunne blant annet nevne at de hadde utfordringer knyttet til bias. De er usikre på hvordan de skal håndtere dette i sine modeller. Forskjeller og urettferdighet i samfunnet vil kunne gjenspeiles i modellene.

Når det gjelder tjenesten for å kunne gi innsikt i arbeidsledige vil dette kunne gjenspeiles i algoritmen, det kommer tydeligere frem. Dette gjelder blant annet borgere med utenlandsk opphav som kan ha vanskeligheter med å få jobb i Norge. NAV ønsket ikke å tillate dette. Leder A sier at*"Det blir på en måte rasisme satt i system. Man innfører bias ved at algoritmen predikerer at man vil kunne få vanskeligheter. To like personer, med samme utdanning, samme*

alder, samme erfaring kan få to ulike resultater på grunn av hudfarge. Man risikerer å gi folk en dårlig start med en gang".

Etterprøvbarhet er et viktig prinsipp når det gjelder utvikling av modeller hos NAV. Som bruker i NAV skal man kunne vite hvorfor en modell kom frem til det resultatet den gjorde. NAV har planer om å publisere algoritmene, de skal forklare algoritmen slik at den blir etterprøvbar for brukerne. Brukerne skal kunne teste algoritmen med sin egen informasjon og det skal bli det samme resultatet som når de kjører den. Her dukker det opp utfordringer ettersom det er maskinlæring, hvor algoritmen lærer etter hver gang den kjører. NAV utdyper at det vil bli vanskelig å forklare hvorfor en konkret gjennomkjøring fikk det resultatet den gjorde. Inntil videre er dette satt på vent da det trengs mer arbeid og bruk av statiske data før en slik løsning kan lanseres.

Skatteetaten
Data som underbygger modellene er nødt til å være av god kvalitet skal modellene fungere. For å være på den sikre siden, blir ikke vedtak fattet i en automatisk beslutningsprosess gjennomført av modellene. Den ansatte er nødt til å gå tilbake til datakilden, før prosseseringen fant sted, for å sikre seg om at funnene stemmer. Leder B forteller at *"Skal man prosessere er det en risiko at man drar feil slutning. Vi har hatt datavarehus i 15 år, og stående regel er at du fatter ikke vedtak eller ekstra skatt ved det du finner. Du går ved datakilden før det er prosessert. I verste fall blir analysen feil, og det er ikke sikkert at noen vil merke det"*.

De forteller videre at det er viktig å ha god datakvalitet for å sikre at ingen blir mistenkt ved uhell. Har man dårlig data får man dårlig resultater, det er derfor viktig at man sjekker flere kilder før man starter kontroll.

Er det feil i datagrunnlaget eller uetiske kombinasjoner, som inkluderer etnisitet eller seksuell legning, kan det bli et galt utgangspunkt. De forteller at i de tidligere modellene som ble utviklet, kom de frem til at menn var dårligere til å betale skatt. Dette kunne ført til at menn ble disproporsjonalt behandlet dårligere. Det viste seg at de måtte jobbe mer med statistikk om inntektsforhold og arbeidsforhold, ikke kjønn. De forteller at det er viktig å ikke dra konklusjoner for tidlig. Leder A forteller at *"Det er viktig å finne den ekte årsaken, som er noe annet enn korrelasjon. Skal du jobbe med dataene og finne ting må du jobbe på riktig måte"*.

SSB
Med tanke på datakvalitet, forteller SSB at endimensjonale stordatakilder er en utfordring. Det gjør at dataen ikke dekker behovet de har i stordata sammenheng.

En annen utfordring er at man er ute etter å samle inn data, men teknologien og metoden gjør det utfordrende. Et eksempel blir gitt der man samler data fra et simkort om hvor en person beveger seg. Man kan ikke være sikker på at det kun er denne personen som har hatt mobiltelefonen med seg hele tiden, og ikke gitt den bort til noen andre. Dette kan medføre at dataene ikke blir korrekte i forhold til hvordan personen beveget seg.

En tredje utfordring er at standarden på dataen kan være rotete. Dataene er ikke nødvendigvis skapt for statistikkformål, og kan være strukturerte på mange ulike måter. De forteller at ofte trenger datakildene mye forarbeid før de kan bli brukbare, slik at man kan trekke konklusjoner ut i fra dem.

En utfordring er representativitet. Ansatt A forteller at *"..når du får disse datakildene eller datasettene, eller datastrømmene. Så kan du ikke være helt sikker på om de er representative med hva du faktisk ønsker å måle. I stordatakilder har man ikke kontroll over designet, du tar det du får. Dette kan medføre at man trenger noen iterasjoner der man lærer seg hva man kan bruke dataen til, og hva det ikke fungerer for"*.

Representativitet, validitet, standardisering og måter å håndtere data med granularitet er store utfordringer i følge SSB.

Datatilsynet
For datakvalitet og bias, kjenner Datatilsynet seg igjen i denne problematikken. Det er et problem med digitale skygger, sier Datatilsynet. At man som elev i skolen kan blir profilert som en skoletaper på grunnlag av data som staten har tilgang på. Slike skygger kan det være vanskelig å vri seg unna. *"Der data aldri slettes får man aldri begynt på nytt"* (Datatilsynet Ansatt A). De peker på personvernlovgivningen og retten til å bli glemt, at data skal kunne slettes, at man skal kunne få en frisk start. De forteller at på nettet i dag kan dette være vanskelig da data følger deg gjennom hele livet.

Etterprøvbarhet, og at ting skal være etterrettelig er krav som Datatilsynet setter. Dette er en utfordring i stordatasammenheng, sier Datatilsynet. At ting skal kunne begrunnes og at de er forutsigbare. De sier blant annet at det er artikler som skal sørge for at beslutninger er rettferdige, gjennomsiktige og etterprøvbare. Når det gjelder stordata ser de en "svart boks-problematikk" når det gjelder disse temaene. Det er vanskelig å forklare hvorfor beslutninger blir som de blir. De presiserer at i vårt lovverk har en krav på å få vite hvorfor ting blir som de blir. *"En har krav på å få en begrunnelse på hvorfor algoritmen kom til den beslutningen den gjorde"* (Ansatt A, Datatilsynet).

Videre er det viktig i stordatasammenheng at beslutninger som treffes skal være rettferdige. Datatilsynet ser på det som urettferdig behandling dersom noe statistisk sett er riktig, men faller uheldig ut for personer. Det kan for eksempel være å belaste brukere mer fordi de ikke bor innenfor et bestemt geografisk område.

5.3.3 Åpne data

Både NAV, Skatteetaten og SSB publiserer mye data som open data. Dette gjøres for å bidra til åpenhet i samfunnet. Bak dette ligger det noen elementer som må på plass for å sørge for tilstrekkelig med personvern.

NAV
NAV legger ut åpne data slik at folk skal kunne forske på dem. Dette kan være masterstudenter, doktorgradsstudenter og forskningsinstitusjoner. Ved å gjøre dette kan man som Leder A uttrykker; *"få hundre flere som forsker på dataene våre enn bare oss selv"*. Innsikten som kommer fra dette er både nyttig for samfunnet og for NAV. De ønsker å hjelpe forskere samtidig som det begrenser antall forespørsler inn til dem. *"Vi har dataen, vi har lagt det ut, dere må bruke det"* (Ansatt A, NAV).

Når det gjelder personvernet så er disse dataene aggregerte. Dersom en som forsker for eksempel skal se på utvikling av arbeidsledighet i en kommune vil en ikke komme ned på individnivå. Det eneste stedet NAV har lagt ut åpne data med det de omtaler som et "en-til-en" forhold er i stillingsdatabasen deres. Her rapporterer bedrifter inn ledige stillinger til NAV, det hentes også inn stillinger fra tjenester som Finn.no.

I stillingsdatabasen kan en gå tilbake i tid for å se hvilke stillinger som lå ute for flere år siden. Her kan man for eksempel se etter etterspørselen på oljeingeniører de siste ti årene, en kan gå inn på den enkelte stillingsannonsen. Her har NAV valgt å ta en etisk beslutning på å fjerne personidentifiserende informasjon på søkere selv om de ikke er lovpålagt å gjøre det. Dette gjelder informasjon som navn, telefonnummer, og i enkelte tilfeller e-postadresser. Ellers kan forskere gjøre tekstanalyser i stillingsannonsene for å for eksempel finne ut hvilke ord en brukte i annonser for ti år siden.

Dersom NAV skal publisere åpne data med statistikk er de nøye med å fjerne felter som kan identifisere individer. De har regler som sier at dersom en celle er færre enn fire så skal det fjernes. Hvis det for eksempel på et sted kun er én arbeidsledig, så skal det ikke komme noe tall. De vurderer også farene for re-identifisering når det gjelder kobling av ulike datasett. NAV ønsker ikke å avsløre personer, og det er der de oppgir at de har brukt mest tid. *"Hvis du har tall om antall personer i Utsira og slår disse sammen med andre data, hva skjer da? Vi skal ikke avsløre personer. Det er der vi har brukt mest tid"* (Ansatt A, NAV).

Skatteetaten
Det er lite data som publiseres åpent, bortsett fra noe aggregert statistikk.

SSB
I forhold til åpne data inngår SSB i ESS, det Europeiske Statistiske Systemet. Det betyr at de følger de retningslinjene som kommer fra EU. I EU har de gjennomført mange år med åpne data prosjekter. Ansatt A forteller at *"..der skjer det veldig mye, og har skjedd veldig mye forskning om å legge ut statistikkdata som er åpne"*. De legger ut aggregert data om for eksempel

befolkningen i Oslo. Regelverket som finnes setter begrensninger for hva de kan legge ut, og de har ansvar for det som legges ut som åpne data.

Når det gjelder publisering, er dette strengt regulert. Det er mange innenfor helseområdet som er interessert i å basere seg på SSB sine data. Et eksempel som nevnes er en studie på medisin, der man vil ønske å ha informasjon som SSB sitter på, om boligforhold eller lignende. De sier at det er strengt å få slått sammen dataen.

SSB kobler sammen data og sørger for at det er anonymiserte filer, til og med når de har slått sammen grupper. Ansatt A sier at "..for å minske risikoen for at enkeltindivider kommer til skade eller identifiseres så finnes det ulike regelverk for hvordan man kan slå sammen grupper". For å sikre personvern er det regelverk, inputkontroll og sikkerhetskontroll, i tillegg har de interne kontroller og regelverk for publisering. Det ser for seg at det samme regelverket vil gjelde for stordata. De jobber kontinuerlig med å sikre at det finnes ulike regelverk for hvordan man kan slå sammen grupper.

Datatilsynet

Når det gjelder bruken av åpne så opplever Datatilsynet at norsk offentlighet er veldig forsiktige. De sier at de per tidspunkt ikke har noen bekymringer innenfor dette temaet, annet en at organisasjoner må tenke på re-identifisering og kobling av ulike datakilder. *"Hvis man slipper ut bostedsdata, de dataene det offentlige legger ut, kan det kobles med andre registre som ligger der ute. Dette kan gi merverdi til disse dataene, og gjør dataene identifiserbare. Så langt så utviser myndighetene stor forsiktighet på dette området"* (Ansatt A, Datatilsynet).

Tabell 6 - Utfordringer knyttet til analyse av stordata

Organisasjon	Anonymisering	Datakvalitet	Åpne data
NAV	Alle fødselsnummer og personnummer byttes ut med en NAV-identitet. Kan ikke spores tilbake til kilden. Sier samtidig at det er en fare for re-identifisering dersom man driller hardt nok ned i dataene. Vurderer støyfilter.	Opptatt av å sørge for at dataene er korrekte. Forskjeller og urettferdigheter i samfunnet vil kunne gjenspeiles i modellene. Utfordrende å håndtere bias. Et mål er at brukere selv skal finne ut av hvorfor en algoritme har kommet frem til et resultat.	NAV legger ut åpne data så folk skal kunne forske på dem. Ønsker å hjelpe forskere samtidig som det begrenser antall forespørsler til dem. Dersom de publiserer åpen statistikk fjernes felter som kan identifisere individer.
Skatteetaten	Personidentifiserende data anonymiseres. Utfordrende med modeller som gjør systematiske feil. Kombinasjoner av opplysninger kan være indentifiserende.	Modellene må være av god kvalitet dersom de skal fungere. For å være på den sikre siden ønsker de ikke at modeller skal ta beslutninger. Viktig så personer ikke blir mistenkt ved uhell. Utfordrende å håndtere bias.	Lite data publiseres åpent, bortsett fra aggregert statistikk.
SSB	Identifiserende elementer fjernes etter det statistiske behovet er oppfylt. Jobber på et overordnet nivå, de benytter seg ikke av data ned på individnivå.	En utfordring med endimensjonale datakilder. Det gjør at dataen ikke dekker behovet de har i stordataprosjektene.	Regelverket setter begrensninger for hva de kan legge ut. De kobler sammen data og sørger for at det er anonymiserte filer, til og med når de har slått sammen grupper.

6. Diskusjon

Gjennom intervjuene har vi kommet frem til en rekke problemstillinger, utfordringer og dilemmaer relatert til personvern og bruk av stordata i offentlig sektor. NAV, Skatteetaten og SSB har i ulik grad forberedt og utført stordata-prosjekter, der NAV har kommet lengst. SSB og Skatteetaten jobber med å utforme og planlegge prosjektene sine, og har reflektert rundt personvern problematikk som oppstår gjennom innsamling, lagring og prosessering av stordata.

Flere begrensninger for stordata kommer gjennom regulative hold, der lovgivning og hjemmelsgrunnlag spiller en stor rolle for hva samles inn og hvordan de kan benyttes. GDPR har gitt et økt fokus på personvern for organisasjonene, der flere jurister arbeider med å tilfredsstille kravene.

I litteraturen er det mange personvernutfordringer og problemstillinger som blir beskrevet fra et overordnet nivå, som ofte gjelder sosiale medier og større teknologibedrifter. Vårt ønske med denne oppgaven er å finne ut hvilke problemstillinger som relaterer seg til norsk offentlig sektor, og de dilemmaene som organisasjonene jobber med.

6.1 Anonymisering

Under våre intervjuer kom vi frem til en rekke personvernutfordringer beskrevet av Skatteetaten, NAV og SSB tilknyttet arbeid med stordata, der en av de mest sentrale er anonymisering av data. Det eksisterer ulike definisjoner av personvern, men det handler først og fremst om retten til å bestemme over egne personopplysninger og retten til privatliv (Datatilsynet, 2017). Gode metoder for anonymisering blir dermed sentralt for å sikre personvern.

NAV, Skatteetaten og SSB forteller oss at de benytter seg av anonymiseringsteknikker for å sørge for at personvern blir opprettholdt i stordata-prosjektene. Leder A hos Skatteetaten forteller at det å anonymisere data tilstrekkelig er en utfordring: *"Det er en utfordring i det å kunne automatisk søke gjennom tekstdokumenter for å fastslå om det finnes identifiserbare personer i dokumentet eller ikke. Kombinasjoner av opplysninger kan være identifiserende"*. Dette bekrefter funnene i litteraturen, og viser at de er bevisste ved problemstillingen.
Gjennom litteratursøket vårt tyder det på at man aldri kan bli helt trygg på at anoymiseringsteknikker for stordata vil kunne garantere anonymitet. Aggregering av data kan benyttes til å gjøre semi-anonym og ikke identifiserbar informasjon til identifiserbar informasjon (Kshetri, 2014). Dette skaper et dilemma, mellom stordata sin intensjon om å samle mest mulig data og koble mest mulig, og hensikten med å sikre personvern.

Fare for re-identifisering
Stordata beskrives i den tradisjonelle 3V-modellen som noe som inneholder data i et stort volum, og med stor variasjon. Det blir derfor naturlig å koble sammen datasett som tilsammen skaper større volum variasjon (Elgendy & Elragal, 2014). Det er klart at stordataanalyse vil tjene på å basere seg på så mye data som mulig, da analysene kan bli enda mer effektive når man har et stort datagrunnlag.

En sentral personvernutfordring, som ble beskrevet av flere av informantene, går ut på anonymisering og faren for re-identifisering. Dette gjenspeiler litteraturen om stordata, der vi finner at det eksisterer en fare for re-identifisering ved at datasett kobles sammen. Selv datasett som er anonymiserte ikke kan identifisere noen på egenhånd, kan nye koblinger med andre datasett utgjøre en risiko for personvern (García Márquez & Lev, 2017; Desouza & Jacob, 2014).

Vi anser dette som en av de større og usikre personvernutfordringene. GDPR definerer en rekke personvernprinsipp, men disse gjelder for personopplysninger, og ikke data som i sin helhet er anonymisert (van Loenen et al., 2016; Mourby et al., 2018). Risikoen for identifisering beskrives av informanten fra Datatilsynet: *"Min sjef pleier å si at det skal bare to datapunkter til for å identifisere han. Det er at han heier på et bestemt fotballag, og at han eier en katt som er av en veldig spesiell rase"*. Vi innser at dette kan skape et dilemma for bruk av stordata, og at det er noe offentlige organisasjoner må være bevisste på skal de sikre innbyggernes personvern.

Cloud Security Alliance beskriver i sin rapport om potensielle sikkerhetsutfordringer ved stordata, at å anonymisere data for analyse ikke er nok for å garantere personvern (Cloud Security Alliance, 2012). En mulighet er å implementere overvåkning av systemet og loggføring, som Skatteetaten har arbeidet med i sin nye plattform Minerva, der de jobber med granulær tilgangskontroll og loggføring. Et mål med denne plattformen er at den blir utviklet med "privacy by design", som betyr at den er designet med utgangspunkt i personvern.
I litteraturen argumenteres det for at data bør anonymiseres tidlig i analysefasen for å hindre at personer kan identifiseres.(García Márquez & Lev, 2017). Dette er et mulig tiltak som kan bedre personvernsikkerheten. NAV forteller at de vurderer å ta i bruk et støyfilter for å anonymisere data til en viss grad. Å legge til "støy" i resultatene er også en metode som beskrives i rapporten fra Cloud Security Alliance for å sikre personvern (Cloud Security Alliance, 2012). Støyfilter kan være en løsning for bedre anonymisering, men spørsmålet blir om det ikke også kan gi en falsk trygghet, og at fremdeles eksisterer en personvernrisiko.

Fra litteraturen og gjennom intervjuer med organisasjonene bekreftes det at selv tilsynelatende anonymiserte data i visse tilfeller kan bryte med personvernprinsippene som er beskrevet av det nye direktivet. Skulle data bli de-anonymisert og personidentifiserende opplysninger bli avslørt gjennom kobling av datasett og analysemetoder, vil det antagelig på nytt falle under direktivets personvernprinsipper. Implikasjonen av dette innebærer at organisasjonene må arbeide med forståelse av risiko ved dataanalyse, kobling av datasett, fare for re-identifisering og juridiske begrensninger.

Balanse brukbarhet og anonymisering
En kan tenke seg at organisasjoner vil ønske å gå for de mest robuste anonymiseringsteknikkene og fjerne ut enhver form for personidentifiserende informasjon. Dette vil imidlertid føre til at datasettene kan risikere å bli mindre brukbare, hevder Mehmood et al. (2016). Dersom et datasett er anonymisert i for stor grad kan det være vanskelig å dra nytte av dataene (Mehmood et al., 2016). Dette kan føre til en utfordring for organisasjonene i fremtiden, da det krever teknisk kompetanse for å vite i hvor stor grad man kan kjøre brukbar analyse på anonymiserte data. En utfordring som går igjen i litteraturen er nettopp det å finne balansegangen mellom tilstrekkelig anonymitet og brukbarhet (Desouza & Jacob, 2014; Mehmood et al., 2016). Dette bekreftes av Leder B for Skatteetaten, som forteller at ved kobling av datasett fra flere ulike kilder vil det være en utfordring å kunne bevise at det finnes identifiserbare personer eller ikke. Han forklarer videre at det ligger begrensninger i de verktøyene som skal løse problemet. Ved å søke igjennom og identifisere om det eksisterer personopplysninger i et datasett, vil resultatet potensielt være villedende og ingen form for garanti for anonymitet selv om den ikke finner noe. Dette er noe organisasjoner bør være bevisste på når de utarbeider sine stordata-prosjekter. Vi mener derfor at det å utvikle verktøy som best kan finne ut risiko for re-identifisering ved bruk av sensitiv data vil være sentralt, eventuelt anskaffelse av verktøy som kan gjøre denne jobben.

Også SSB forteller at når de kobler sammen data sørger de for at det er anonymiserte filer; *"for å minske risikoen for at enkeltindivider kommer til skade eller identifiseres finnes det ulike regelverk for hvordan man kan slå sammen grupper"* (Ansatt A, SSB). Det er med andre ord også juridiske begrensninger som må tas hensyn til, ikke kun tekniske.

Det blir en avveining for organisasjonene hvor mye ressurser de skal benytte på å arbeide med å prøve å sikre anonymisering. Det er fremdeles tidlig i prosjektene, og de arbeider med å ta i bruk stordata og bevise hvilke fordeler det kan bringe til organisasjonene. I tråd med Mai (2016) sitt argument for en ny modell for personvern og stordata, fremstilles en påstand om at de største personvernutfordringene kommer fra de nye kunnskapene, innsiktene og realitetene som skapes gjennom data som er innsamlet. Det vi vil anse som den største utfordringen vi har funnet gjennom studien er nettopp personvernutfordringer knyttet til analyse, som datafikasjon modellen argumenterer for, og bør være et fokus for organisasjonene etterhvert som stordata blir brukt mer og mer.

6.2 Profilering og bias

Tilstrekkelig datakvalitet har vi fått inntrykk av at vil være særdeles viktig for norsk offentlig sektor. Datakvalitet omfavner også den fjerde V-en i stordata definisjonen, veracity - troverdighet (Elgendy & Elragal, 2014). I stordataanalyse er det viktig at det er korrekt og troverdig informasjon som samles inn slik at resultatene ikke blir en tilnærming. Resultatene av en stordata analyse skal være en representasjon av virkeligheten (Maciejewski, 2016).

Dersom datakvalitet i stordata-prosjektene er av dårlig kvalitet blir resultatet mindre representativt. Konsekvenser av dette er at brukere av systemene kan risikere å bli 'stemplet',

eller satt i feil bås dersom dataene skal bli brukt til å ta beslutninger (Rogge et al., 2017). Norsk offentlig sektor bør være klar over at det kan være svært vanskelig å komme seg ut av et slikt resultat. Fra et personvernperspektiv er utfordringen bias og profilering. Datatilsynet nevner dette med digitale haler som et eksempel. Hvis en person i gjennom en stordataalgoritme blir stemplet som en skoletaper kan dette følge han, eller henne, gjennom hele livet.

SSB sier at man ikke kan være helt sikker på om dataene er representative med hva man faktisk ønsker å måle. Her er det en utfordring med endimensjonale stordatakilder. Det blir vanskelig å tyde dataene, slik at de ikke egnes for statistikkformål. Kaisler et al. (2013) argumenterer for at informasjon fra modeller må være nøyaktig dersom man skal ta beslutninger basert på stordata. De mener at organisasjonen må spørre seg selv hva som blir et riktig estimat, det kommer an på organisasjonen som tar i bruk slike stordataanalyser (Kaisler et al., 2013). Skatteetaten tar dette alvorlig, og vil ikke at la algoritmer ta automatiske avgjørelser basert på dataen de har samlet inn. De går tilbake til kilden før de tar en avgjørelse.

At data skal være riktig ser Datatilsynet på som et grunnleggende personvernprinsipp. Det er også et punkt i den nye personvernforordningen, GDPR. I sin veileder for organisasjoner om det nye regelverket skriver de at "personopplysninger som behandles skal være korrekte". At Skatteetaten ikke velger å stole på modellene, viser at de følger personvernprinsippene til GDPR. De er klar over at det er en risiko for at datakvalitet kan påvirke resultatet av stordataanalyse. Det er positivt at de tar hensyn til dette, spesielt når det er snakk om prosjekter som kan påvirke folk i stor grad (Rogge et al., 2017).

Når det gjelder løsninger på denne problemstillingen har NAV sett på hvordan de kan gjøre modellene sine etterprøvbare. Hos NAV skal brukere selv ha mulighet til å se hvorfor systemet har kommet frem til et resultat. En slik løsning kan bidra til å luke ut feil i modeller, og kan være en interessant løsning på å kontrollere riktighet. For den offentlige sektoren kan etterprøvbarhet også bidra til gjennomsiktighet, som er beskrevet som et personvernprinsipp (Datatilsynet, 2017). Dersom brukere får innsikt i hvordan tjenestene deres fungerer kan dette bidra med å gi tillit. For både NAV, SSB og Skatteetaten er tillit viktig for å få gjennomslag i stordata-prosjektene, det er ikke slik at de ønsker å teste grensene når det gjelder disse områdene. Likevel kan det være vanskelig å formidle hvordan akkurat 'den' algoritmen kom frem til det svaret den gjorde.

Etterprøvbarhet vil følge kravene til Datatilsynet om riktighet. De vil at brukere skal få vite grunnlaget av et resultat basert på stordataanalyser. De sier nemlig at de ikke er uvanlig med diskriminering når det gjelder slike tjenester. De kommer med et eksempel hvor sier at det ofte er slik at personer som bor i enkelte områder av landet får flere goder enn andre. Skatteetaten forteller at det kan oppstå systematiske feil i modellene som er viktig å oppdage, som gjør etterprøvbarhet viktig.

Bias er på mange måter en vanskelig problemstilling. Modellene er en tilnærming av virkeligheten. Dette vil si at denne tilnærmingen blir en gjenspeiling av samfunnet i norsk offentlig sektors storadatasatsning. Dersom det ligger diskriminering i datasettene vil resultatene av analysen kunne gjenspeile dette. Modellene har ingen innebygd filter for å håndtere

diskriminering. I en jobbsøketjeneste kan en se for seg to personer med akkurat samme utdanning og arbeidserfaring. De vil kunne få et helt ulikt resultat basert på hvilket land de kommer fra. NAV kaller dette problemet for 'rasisme satt i system', hvor det ikke tas hensyn til individet. Det er på mange måter snakk om et etisk problem. For NAV sin del har de besluttet at dette ikke er faktorer de ønsker å ha i sine løsninger.

6.3 Formålsbegrensning, dataminimering og samtykke

Vi vil i denne delen argumentere for at samtykke, formålsbegrensning og dataminimering henger alle sammen, og vil spille en begrensende rolle for hvordan data kan benyttes i stordata-prosjektene. At disse prinsippene følges, vil vi påstå er viktig med tanke på risikoen for re-identifisering, profilering og bias som vi har beskrevet tidligere.

Dataminimering
Innsamling er en av de tre delene i stordata livssyklusen, der dataminimering vil spille en sentral rolle for å kunne sørge at personvern ivaretas. Dataminimering har fått nytt fokus ved innføringen av GDPR. Datatilsynet beskriver på sin veileder om innføring av det nye regelverket, at dataminimering innebærer å «begrense innsamlede personopplysninger til det som er nødvendig for å realisere innsamlingsformålet» (Datatilsynet, 2017). Informanten fra Datatilsynet beskriver dataminimering og stordata som en motsetning, fordi stordata handler om å samle inn så mye data som mulig. Dette bekreftes av 3V modellen, der volum er er en av de 3 V-ene i modellen som beskriver stordata (Chen, Lao & Liui 2014). Vi ser at begrensningene dataminimering setter blir en utfordring, når stordataanalyse ofte handler om å finne nye innsikter som ikke eksisterte tidligere (Kshetri, 2014).

Litteraturen poengterer at når individer utgir informasjon, gjøres det innenfor et sett med personvernregler (Hilbert, 2016). Skulle data innsamles og brukes i stordataanalyse for å skaffe innsikter som ikke var samtykket til, er dette et regelbrudd og kan føre til tap av omdømme for organisasjonen. Skatteetaten nevner tap av omdømme som en konsekvens skulle data på noen måter misbrukes til andre formål enn det opprinnelig var beregnet for.

Det kan være utfordrende å forklare alle mulige metoder som data kan brukes på ved innsamling, fordi stordataanalyse er skapt for å hente ut gjemt informasjon, finne korrelasjoner mellom datasett og finne uforutsigbare slutninger (Mantelero & Vaciago, 2015). Under intervjuene med NAV kommer det frem at de ikke har fokusert på dataminimering før lovverket ble aktuelt, og at de sørger nå å følge loven for utviklingen av stordatatjenestene med hjelp av Datatilsynet. De forteller videre at de har gjennomført stordataprosjekt som er helt i tråd med det nye direktivet. De forteller at *"Vi ønsker tillit fra brukerne, og vil ikke presse grenser"* (Leder A, NAV). Vi mener at det vil være sentralt for organisasjonene at de fortsetter å samarbeide med Datatilsynet, og sikrer videre tillit fra brukere.

Da NAV har kommet lengre med stordata-prosjektene sine, er det ikke overraskende at de har hatt enda større fokus på å følge GDPR enn Skatteetaten. En ansatt bemerker at tjenestene kunne vært bedre dersom de hadde hatt et større datasett, som tilsier at prinsippet om dataminimering

kan være en hindring for stordata bruk, og kan sees på som en balansegang mellom personvern og utbytte av stordata.

Man kan påstå at dataminimering ikke er den største personvernutfordringen for NAV og Skatteetaten for øyeblikket, da det heller er bruken av data i ettertid; formålsbegrensing, som blir juridisk og etisk utfordrende for organisasjonene. Organisasjonene gir for øyeblikket ikke inntrykk av å samle inn mye data spesifikk for stordatabruk. Det er heller bruk av eksisterende data som benyttes til analyseformål, som blir innsamlet med hjemmelsgrunnlag.
Vi ser dermed at dataminimering og formålsbegrensning kan bli sett på som tett knyttet sammen. Der dataminimering setter begrensninger for hva som kan samles inn til stordataformål, vil formålsbegrensning sette begrensninger for hvordan data som har blitt samlet inn til andre formål benyttes i stordatasammenheng.

SSB forteller at det ligger strenge retningslinjer for innsamling og bruk av data, der også juristene spiller en rolle for hvordan opplysninger kan benyttes. SSB samler kun data hvis det er statistikkformål til det, her må det også ligge et formål dersom de skal benytte seg av den.

Ved å benytte seg av jurister og eksperter fra Datatilsynet, tar Skatteetaten, NAV og SSB dataminimering alvorlig. Både Skatteetaten og NAV gir uttrykk for at dataminimering setter begrensninger for hvordan offentlig sektor kan utføre stordata-prosjekter. Personvernprinsippet om dataminimering setter stopper for utfordringer som ligger ved å samle inn for mye data om befolkningen. Samtidig fører dette til at organisasjonene får et mindre brukbart datasett.

Forsmålsbegrensing
Formålsbegrensing er et personvernprinsipp beskrevet av GDPR og Datatilsynet, der virksomheter må forsikre seg om at data brukes til det originale innsamlingsformålet (Datatilsynet, 2017).

Organisasjonene sier selv at dette begrenser måten de kan bruke stordata på. Organisasjonene har mye data internt, der potensialet ikke blir utnyttet. Samtidig er dette et viktig personvernprinsipp den offentlige sektor bør følge. Da offentlige organisasjoner har hjemmel til å samle inn sensitive opplysninger er det viktig at dataene brukes på en forsvarlig måte. Selv om dataene er anonymiserte har forskning vist at det ikke skal mye til før en bryter sentrale personvernprinsipper. Likevel er det slik at disse prinsippene er et relativt nytt konsept for det offentlige i sammenheng med stordata-prosjekter. Dette er noe de også gir inntrykk av, da det er knyttet stor usikkerhet til hvordan data kan benyttes. NAV sier selv at det er flere uavklarte elementer når det gjelder deres prosjekter. De som jobber med den tekniske tilnærmingen av stordata har måttet spørre seg selv om hvilke data de har lagret og kan benytte seg av.

Videre er det viktig å få frem at organisasjonene ikke ønsker å teste grensene. Alt i alt handler det jo om å kunne tilby tjenester for brukerne deres. Data blir ikke samlet inn kun for stordatabruk, de blir samlet inn for egne formål. Bruken av eksterne kilder er også viktig her. Når det gjelder bruken av for eksempel sosiale medier til stordataformål utrykker de forsiktighet. Her vil det i enkelte tilfeller kun være aktuelt ved kontrollformål.

Det er også slik at den nye lovgivningen har gjort at organisasjonene i det hele tatt har fått et større fokus på denne problemstillingen. Skatteetaten beskriver at den nye personvernforordningen har satt fokus på hva de har hjemmelsgrunnlag for til videre bruk. Leder B forteller at *"En gang ble det sagt, det var før personvernordningen kom, så brukte man begrepet alt vi vet er alt vi gjør. Det begrepet bruker vi ikke lenger"*.

Samtykke
Når det kommer til samtykker behøver ikke organisasjonene be om dette dersom formålet med stordata-prosjektene er relatert til deres primæroppgaver. Dersom en tar Skatteetaten som eksempel, så betyr dette at de får bruke data som er relatert til skatteformål uten å be om et samtykke. Dette har de et hjemmelsgrunnlag for. Videre blir juridiske beslutninger tatt i organisasjonene der det ligger usikkerheter i hva de kan bruke. Dersom en skal følge GDPR må organisasjonene forsikre seg om at bruken av data stemmer med det opprinnelige innsamlingsformålet. Det vil altså bli aktuelt for organisasjonene å ta i bruk samtykker dersom de går over disse retningslinjene.

Videre kan man stille seg spørsmål om hvordan et slikt samtykke bør se ut. Det får som faktisk leser hva samtykke innebærer (Tan & Pivot, 2015). Det kan være utfordrende for en forbruker å forstå hva han eller henne samtykker til i stordatasammenheng (Moorthy et al., 2015). Datatilsynet sier at folk har en tendens til å godta samtykker uten å egentlig tenke over hvilke følger det får. Brukere er lei av lange stiler om hvordan applikasjoner tar i bruk dataen deres. Her er det viktig at de offentlige får frem budskapet på en kortfattet og brukervennlig måte. Fra Datatilsynets side anbefales det å ta i bruk ikoner og illustrasjoner for å forenkle denne prosessen for brukerne. Dette kan også være et viktig grep for å bygge tillit.

6.3 Lagring og tilgangskontroll

Bruken av skylagring
Ved bruken av skylagring er det vanskelig å garantere for at dataene er sikre (Hashem et al., 2015). Det er flere faktorer som spiller inn. Først og fremst så er det slik at lover og regler ikke er like i alle land. En organisasjon kan av den grunn ikke garantere for at leverandøren følger de samme lovene og reglene som dem selv. Tankard (2012) trekker blant annet frem eksempler på at overvåkning i noen land er tillatt ved spesielle tilfeller (Tankard, 2012). Datatilsynet anbefaler å ta i bruk leverandører som følger europeisk lovgivning. Grunnen til dette er at en kan garantere for en grad av sikkerhet når det gjelder bruken av skylagring og eksterne leverandører. En har også tilfeller hvor større organisasjoner som tilbyr skylagring har opprettet datasentre for å berolige kunder. Dersom organisasjoner skal ta i bruk skylagring må de stille seg spørsmål om loven i landet som leverandører holder til i kan gi tilstrekkelig med beskyttelse.

En annen faktorer er mangel på kontroll over dataene som er lagret i skyen. At dataene ligger i noens andre hender, gjør at det faller ned på tillit og integritet til leverandøren en benytter. Organisasjoner bør stille krav til at data kun skal endres av autoriserte parter eller dataeiere slik

at en kan forhindre misbruk (Hashem et al., 2015). Tilgjengelighet er også et viktig begrep. Hvis man skal ta i bruk leverandører stiller det krav til at plattformen er tilgjengelig til enhver tid. Sett fra et personvernperspektiv kan det være vanskelig å garantere at slike tjenester vil være tilgjengelige dersom de skulle være involvert i et datainnbrudd eller hackerangrep (Zissis & Lekkas, 2012).

Denne usikkerheten gjør at skylagring er uaktuelt for norsk offentlig sektor når det gjelder anvendelsen av stordata. Det er for risikabelt å la data ligge lagret hos eksterne leverandører da man har for liten grad av kontroll. En kan også stille seg spørsmål om tilgangskontrollen tilfredsstiller den graden av sikkerhet som kreves av det offentlige. Data lagres derfor lokalt hos både NAV, Skatteetaten og SSB. Organisasjonene gav inntrykk av at lagring var et tema de ønsket å snakke lite om. Ansatte hos NAV så på det som en risiko at vi i det hele tatt spurte dem om temaer knyttet til lagring. Spesielt interessant var det hvordan ansatte i NAV hadde delte meninger om skylagring og fremtiden. Ansatte som jobbet med de tekniske aspektene ved stordata uttrykte stor bekymring knyttet til temaet. Ansatte på fagsiden var heller opptatt av mulighetene som stordata vil gi. En ansatt hadde også en formening om at stordata kunne være enda tryggere i skyen.

Dersom en skal se på fremtiden er det vanskelig å se for seg at skylagring blir uaktuelt. Spesielt i sammenheng med stordata. Ettersom stadig større mengder med data genereres og samles inn kreves det alternative metoder for å følge tritt med utviklingen (Chen et al., 2014). Skylagring spås som et lovende alternativ til tradisjonell lagring. Fordelene med skylagring er at en slipper kostnader som vanligvis er knyttet til fysiske lagringsplattformer (Leavitt, 2013). Dersom norske offentlige organisasjoner hadde valgt å ta det skylagring i bruk ville det sørget for en mer effektiv plattform. De slipper å stadig oppgradere infrastrukturen som følger av utvikling i stordata-prosjektene deres. Samtidig vil dette gjøre at organisasjonene kan fokusere på deres kjerneområder, da en tredjepart ville håndtert utfordringer knyttet til lagring. Samtidig er det viktig å få frem at tradisjonelle databasesystemer i mindre grad kan håndtere ustrukturerte data som lyd og bilder (Chen et al., 2014).

Selv om det i dag er uaktuelt for både NAV, SSB og Skatteetaten å ta i bruk skylagring, utelukker ikke at det vil bli aktuelt å ta det i bruk på et senere tidspunkt. Hvis norsk offentlig sektor skal velge å bruke skylagring er det viktig med gode databehandleravtaler. Det bør tydeliggjøres hvem som sitter med ansvaret. Det bør velges en leverandør med god omtale. Videre bør de etter råd fra Datatilsynet ta i bruk europeiske leverandører, eller organisasjoner med datasentre i Europa. Den norske offentlige sektoren kan få fordeler av å ta til seg kunnskap fra andre land hva de tenker om skylagring.

Metoder for tilgangsstyring

For å sikret at dataene ikke havner hos "ondsinnede" vektlegges tilgangskontroll.
En finner ikke nødvendigvis at tilgangskontroll er en typisk stordata utfordring i litteraturen, men vi fikk lære at det innenfor den offentlige sektoren er et viktig tema. Den offentlige sektoren sitter på store mengder med personidentifiserende data. Det samles inn data om mennesker fra fødselen til døden. Dette fører til at det offentlige er et større offer en private organisasjoner når det gjelder hackerangrep og datainnbrudd. NAV sier selv at de sitter med enorme mengder av

data på den norske befolkningen. Samtidig er det ikke uvanlig at ansatte organisasjonen selv gjør tillitsbrudd. Datatilsynet sier at de vet om hendelser der ansatte har vært nysgjerrige. De kan fortelle at det er kjente personer som er spesielt utsatt i disse tilfellene.

Det er viktig at kun de som har behov for det skal ha tilgang til sensitiv data. Det er flere måter å styre tilgang på, hvor rollestyrte metoder ofte er benyttet. Her må en være oppmerksom på at det kan gis for brede tilganger. Det er utfordrende for organisasjonene å sette riktige tilganger, spesielt når en driver med stordataanalyser. Skatteetaten sier at de har gitt for brede tilganger til ansatte som jobber med modellering av stordata. Dette bekrefter også forskning sier om tilgangskontroller. Definering av roller kan være både tidkrevende og vanskelig, spesielt når ansatte har ulike tjenestebehov (Miller & Mork, 2013). Dette vil si at flere har sett på alternative måter å håndtere dette på, blant annet en innholdsbasert tilgangsbegrensning. Her kan organisasjonene styre tilgang basert på hvilket innhold dataene har. Allikevel er ikke denne metoden like robust som rollebaserte metoder er, hvor den enkeltes behov i større grad vil bli analysert (Zeng et al., 2013).

Det er også slik at enkelte ansatte behøver bredere tilganger for å kunne gjøre analyser. Disse er det imidlertid ikke mange av i organisasjonene vi intervjuet. Hos NAV har man blant annet analytikere som kan se på data ned på individnivå, for å så å løfte dem opp på et mellomlag.

Det er også slik at det kan jukses med tilgangsstyringene. Skatteetaten utelukker ikke at ansatte benytter seg av roller som ikke samsvarer med deres arbeidsområde. En kan på mange måter si at tillit står sterkt når det gjelder tilgangsbegrensning i den norske offentlige sektoren. Organisasjonene stoler på at deres ansatte ikke svekker tilliten deres ved å gjøre ulovlige handlinger. Selv sier de at ansatte undertegner taushetserklæringer som vil hindre ansatte å spre sensitive opplysninger.

Det skal også sies at organisasjonene vi har tatt utgangspunkt i har svært mange ansatte, og at det kan være vanskelig å holde styr på tilgangskontroller. Et effektivt tillegg til en rollestyrt inndeling er å ta i bruk loggføring, sier Datatilsynet. De ønsker at alle tar i bruk logging der det finnes personidentifiserende informasjon. De sier at det ikke skal ikke være slik at naboer kan søke deg opp på arbeidsplassen. Det kan ligge mindre grad av 'snoking' i sensitive opplysninger dersom ansatte er klar over at handlinger de gjør blir logget.

Videre finner vi at det å definere sensitive data i denne sammenheng også er utfordrende. Organisasjonene definerer 'alt' som sensitive data, men her har de også ansvar for personer hvor det kreves mer varsomhet. Dette er gjelder blant annet informasjon om personer som lever på beskyttede adresser. Skatteetaten opererer her med koding, for å beskrive i hvor stor grad dataene er sensitive. NAV opplever at det å ta i bruk slik personidentifiserende informasjon som problematisk, og tar i bruk et personvernombud der det er behov for det.

6.5 Implikasjoner for praksis

I praksis vil vi argumentere for at organisasjonene må jobbe med å kunne identifisere potensiell risiko ved sammensetning av ulike datasett, og finne ut hvordan dette kan misbrukes. Vi blir fortalt at data deles gjennom organisasjonene, blant annet Skatteetaten som benytter seg av data fra NAV. SSB har en rekke data som benyttes til statistikkformål. Brukes denne dataen til stordataanalyse vil det være nødvendig for organisasjonene å ta en risikovurdering. Et viktig poeng er bevisstheten over en underliggende fare for at re-identifisering kan forekomme. Informanten for SSB gir et eksempel i en forskningsstudie for medisin at det ønskes at SSB deler informasjon, slik at det kan sammenslås med statistikk om boligforhold og lignende. Gjennom strenge reguleringer hindrer regelverk at visse sensitive data kan bli slått sammen. Man kan spekulere i hvordan data om medisinsk bruk og boligforhold kunne blitt utnyttet av arbeidsgivere og forsikringsselskap på kyniske måter, skulle denne type data komme på avveie. Til syvende og sist argumenterer vi for at det er integriteten til ansatte som jobber som sørger for at stordata benyttes på en måte som minimerer risiko for re-identifisering. Det innebærer å bygge algoritmer og modeller innenfor dataanalyse som ikke prøver å kombinere data på en måte som kan skape personidentifiserende resultater. Det finnes en rekke metoder som detaljert beskriver slike teknikker for anonymisering som denne studien ikke går inn på.

Vi vil også argumentere for at det må være klart definerte rammebetingelser for hvordan data benyttes i stordatasammenheng, og organisasjoner må ha et klart definert formål med analysene. Har de ikke dette, ser vi at det utfordrende å finne ut hvordan de på best mulig måte kan minimere innsamlingen av data. Organisasjonene var ikke helt forberedt på GDPR da vi intervjuet dem, men de jobbet å sørge for å følge personvernprinsippene. De gir uttrykk for at disse prinsippene kan være til hinder for full utnyttelse av stordata. Vi mener at de må ta disse prinsippene på alvor, og tydeliggjøre hvordan de kan komme i konflikt med stordataanalyse i den fremtidige utviklingen av prosjektene deres.

Datatilsynet forteller at de opplever misforståelser når det gjelder anonymisering av data, der man tror at man har anonymisert det, men i virkeligheten er det kun pseudo-anonymisert. Vi har inntrykk av at de offentlige organisasjonene vi har intervjuet tar anonymisering og faren for re-identifisering alvorlig. Konsekvensen av dette innebærer at man også må sørge for å ha god tilgangskontroll, og sikre sensitive data på måter slik at de ikke lekker ut. Her ligger det en viss risiko for at ansatte får for brede tilganger. Det kan være vanskelig å gi riktig grad av tilgangskontroll til analytikere.

Videre mener vi at norsk offentlig sektor bør ta hensyn til at det kan ligge bias i datamodellene som ligger i grunn for dataanalysen. Vi mener at modeller bør formes med dette i tankene. Det vil ikke nødvendigvis være enkelt å bevise at en er lovlydig dersom tjenestene sier noe annet. Å gi borgere innsikt i hvordan disse løsningene fungerer i praksis vil være en god ide for å hindre forskjellsbehandling. Samtidig ser vi at det kan være med på å øke borgeres grad av tillit til det offentlige, som er viktig for å få gjennomslag i stordata-prosjekter. At det i dag ikke blir brukt algoritmer til å ta endelige beslutninger er betryggende. Vårt inntrykk er at organisasjonene har en vei å gå når det gjelder disse utfordringene.

Når det gjelder bruken av skylagring er det ikke overraskende at det blir sett på som en risiko. Ettersom de fleste tegn tyder på at skylagring vil være fremtiden når det gjelder stordata er det viktig at norsk offentlig sektor, etter råd fra Datatilsynet, ser på hvordan en kan lage gode databehandleravtaler.

7. Konklusjon

I denne studien har vi undersøkt NAV, Skatteetaten og SSB sin satsing på stordata, med forskningsspørsmålet "Hvilke personvernutfordringer oppstår ved bruk av stordata i norsk offentlig sektor?". Vi har sett at deres bruk av stordata er i et tidlig stadie, og at det fremdeles ligger en del arbeid i å utforme prosjektene og avdekke potensielle personvernutfordringer. For å besvare forskningsspørsmålet har vi gjennomført 14 semi-strukturerte intervjuer, med informanter fra NAV, Skatteetaten og SSB. Alle informantene har hatt tilknytning til stordata-prosjektene i form av ledelse, planlegging eller utvikling. Datatilsynet har som tilsynsorgan også bidratt med sine synspunkter rundt vårt forskningsspørsmål. Vi har utarbeidet en rekke punkter der vi beskriver det vi ser på som de mest sentrale personvernutfordringene som oppstår:

1. Den første utfordringen er knyttet til anonymisering av personopplysninger og sensitive data for bruk i stordataanalyse. Det eksisterer en fare for re-identifisering ved kobling av forskjellige datasett, som er gjeldende selv om informasjonen er pseudo-anonymisert eller anonymisert.
2. Den andre utfordringen er å sikre gode modeller for analyse og at datakvaliteten er representativ, som minsker sjansen for profilering og diskriminering.
3. Den tredje utfordringen gjelder personvernprinsippene beskrevet av GDPR om dataminimalisering, formålsbegrensning og samtykke**.** Disse er ofte i konflikt med full utnyttelse av stordataanalyse. Det oppstår utfordringer knyttet til etterlevelse av disse prinsippene.
4. Den siste utfordringen går ut på å sikre personopplysninger og sensitive data for potensielt misbruk. Det ligger utfordringer i tilgangskontroll og sikker lagring, som for øyeblikket hindrer bruk av skylagring.

Etterhvert som stordata blir tatt i bruk i stadig større grad i offentlig sektor, er det sentralt at personvernutfordringene avdekket i denne studien blir tatt alvorlig, og at det eksisterer en generell bevissthet rundt de. Skal man høste gevinstene av stordata, er det viktig at man også er klar over hvilken risiko dette innebærer for personvern. Offentlig sektor behandler mye sensitiv informasjon som ikke må komme på avveie eller misbrukes. Gode metoder for anonymisering av personopplysninger og sensitiv data, etterprøvbarhet i modellene, regulative begrensninger, bevissthet rundt personvernprinsippene, sikker lagring og tilgangskontroll, samt åpenhet rundt bruk av stordata vil være viktig for å sikre personvern og innbyggernes tillit til offentlig sektors bruk av stordata.

7.1 Studiens bidrag

I denne studien har vi sett hvordan tre ulike offentlige organisasjoner jobber med stordata, og hva de ser på som de største personvernutfordringene. Stordata blir stadig mer relevant, der privat sektor har utnyttet mulighetene lenge, begynner nå også offentlig sektor å se gevinstene for effektivisering og forbedring av tjenester. Vi anser at bruken av stordata i norsk offentlig sektor er på et tidlig stadie, som bekreftes av informantene i studien. Gjennom vårt litteratursøk ser vi at mange av utfordringene som blir beskrevet av informantene ikke kun er relatert til offentlig sektor, men generelle for bruk av stordata. Studien viser at offentlig sektor har spesielle hjemler for innsamling og bruk av data, der det behandles mye personopplysninger og sensitive data om innbyggere. Det er tydelig at offentlige organisasjoner tar personvern alvorlig, og har store juridiske avdelinger og samarbeid med Datatilsynet for å sikre seg at de ikke bryter noen regler. Denne studien bidrar til å bevisstgjøre organisasjoner ved de potensielle farene og usikkerhetene knyttet til stordataanalyse, hvilke personvernutfordringer som finnes, og hvilke konsekvenser manglende personvern kan ha.

Med tanke på innføringen av GDPR og de personvernprinsippene som nå offentlige organisasjoner må ta hensyn til, kan denne studien sees på som spesielt relevant for tidspunktet den publiseres. Fra et akademisk perspektiv, kan studien være et illustrativ case som kan benyttes i undervisningssammenheng, som viser hvilke personvernutfordringer som er de mest sentrale ved bruk av stordata i norsk offentlig sektor. .

7.2 Videre forskning

Denne studien bidrar til en utvidet forståelse av norsk offentlig sektors bruk av stordata, og danner et videre grunnlag for videre forskning:

- Studien legger grunnlag for flere kvalitative og kvantitative studier som utforsker personvernutfordringer knyttet til stordata i offentlig sektor.
- Studien lager grunnlag for studier som sammenligner personvernutfordringer og stordata i offentlig og privat sektor.
- Tekniske utfordringer relatert til stordataanalyse som kan være en trussel for personvern som har ikke blitt utforsket i denne studien. En videre undersøkelse kan ta de tekniske personutfordringene ved lagring, analyse og innsamling av data.
- Modenhet i stordatabruk er ikke undersøkt i denne studien. En undersøkelse som tok for seg modenheten i stordataprosjektene kunne undersøkt:
    - Hvor personvernutfordringer oppstår for hvor velutviklet bruken av stordata er.
    - Hvilke personvernutfordringer som forsvinner ved høy modenhet for bruk av stordata.
- Forskning på norske offentlige organisasjoner, som har kommet enda lengre i bruk av stordata kan bidra til en bredere forståelse av temaet.

7.3 Studiens begrensninger:

Her viser vi hvilke begrensninger som kan ha påvirket resultatet:

- Kun en person var tilgjengelig fra SSB. Dette gjør at informantene kom litt skjevt ut, og har gjort at vi har lagt mer vekt på funnene fra NAV og Skatteetaten. Fra Datatilsynet fikk vi også kun en informant. Hadde vi hatt flere, kunne vi bedret den interne validiteten i studien.
- Studien er begrenset til tre organisasjoner i norsk offentlig sektor som jobber med stordata. Det er flere som gjør det og har kommet lengre, som Tolletaten, som ønsket ikke å delta. Flere organisasjoner og flere intervjuer kunne gjort resultatene mer generaliserbare, og dermed bedret ekstern validitet.
- Ut i fra informantenes beskrivelse om data som benyttes, kan man stille spørsmål hvorvidt deres bruk av data kan defineres som stordata, med tanke på volum, variasjon og hastighet. Det er ingen definert teoretisk grense for hva som defineres som stordata. I fremtiden kan stordata defineres annerledes enn det som ligger til grunn akkurat nå.
- Det lå litt i usikkerhet i hva Skatteetaten definerte som stordata, og om det de holdt på med kunne beskrives som stordata. Noen var mer tydelige enn andre på at de brukte stordata. Dette gjorde at informantene til tider svarte på et mer generelt grunnlag med tanke på potensielle utfordringer, ut ifra hva de selv definerte som Skatteetatens stordatabruk.
- Prosjektene NAV arbeidet med, er fremdeles i startgropen, der de enda arbeider med å bevise at disse prosjektene er fremtiden for NAV. Skal de fullt ut kjøre stordata-prosjektene basert på data fra innbyggerne i stor skala, kan det potensielt dukke opp nye personvernutfordringer som vi ikke har avdekket. Dette gjør at de uttaler seg på mer generelt grunnlag.
- Ledere og personer høyt oppe i stilling har deltatt i denne studien. Vi har derfor ikke fått flere synspunkter fra de som jobber mer teknisk med å utvikle prosjektene.
- Organisasjonene jobber fremdeles med å implementere GDPR. Siden deres bruk av stordata er i en tidlig fase, vil det være en del arbeid de fremdeles har med å oppdage utfordringer knyttet til personvernprinsippene som GDPR definerer.

8. Referanser

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around procedural privacy protections. *Communications of the ACM, 57*(11), 31-33. doi:10.1145/2668897

Bernard, T. S., Hsu, T., Perlroth, N., & Lieber, R. (2017). Equifax Says Cyberattack May Have Affected 143 Million in the U.S. *The New York Times*. Retrieved from https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html

Bertino, E. (2013). *Big Data -- Opportunities and Challenges Panel Position Paper*. Paper presented at the 2013 IEEE 37th Annual Computer Software and Applications Conference.

Bhimani, A. (2015). Exploring big data's strategic consequences. *Journal of Information Technology, 30*(1), 66-69. doi:10.1057/jit.2014.29

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society, 15*(5), 662-679. doi:10.1080/1369118x.2012.678878

Cavanillas, J. M., Curry, E., & Wahlster, W. (2016). *New Horizons for a Data-Driven Economy*. Madrid, Spain.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications, 19*(2), 171-209. doi:10.1007/s11036-013-0489-0

Cloud Security Alliance. (2012). Top ten big data security and privacy challenges. Retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf

Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks: SAGE Publications.

Cumbley, R., & Church, P. (2013). Is "Big Data" creepy? *Computer Law & Security Review, 29*(5), 601-609. doi:10.1016/j.clsr.2013.07.007

Daniell, K. A., Morton, A., & Ríos Insua, D. (2015). Policy analysis and policy analytics. *Annals of Operations Research, 236*(1), 1-13. doi:10.1007/s10479-015-1902-9

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., . . . Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56-63. doi:10.1145/2643132

Datatilsynet. (2016). Hva er personvern? Retrieved from https://www.datatilsynet.no/om-personvern/hva-er-personvern/

Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/

Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7769

Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7770
Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7771
Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7772
Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7773
Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7775
Datatilsynet. (2017, Aug 24, 2017). Grunnleggende personvernprinsipper etter nytt regelverk. Retrieved from https://www.datatilsynet.no/regelverk-og-skjema/veiledere/grunnleggende-personvernprinsipper-etter-nytt-regelverk/?id=7776
Datatilsynet. (2018). Datatilsynet vil møte Tolletaten om bruken av amerikansk IT-verktøy. Retrieved from https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-2018/datatilsynet-vil-mote-tolletaten-om-palantir/
Desouza, K. C., & Jacob, B. (2014). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society, 49*(7), 1043-1064. doi:10.1177/0095399714555751
Elegendy, N. & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. Lecture notes in computer science, 214-227. DOI: 10.1007/978-3-319-08976-8_16.
Fola, M., & Vania, S. (2016). Data intelligence for local government. *Policy Studies Organization*, 21.
Gamage, P. (2016). New development: Leveraging 'big data' analytics in the public sector. *Public Money & Management, 36*(5), 385-390. doi:10.1080/09540962.2016.1194087
Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007
Gjersdal, A. (2018). Metadata. *SNL.* Retrieved from https://snl.no/metadata

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013). *A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case*. Paper presented at the 2013 IEEE International Congress on Big Data.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khanb, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 18. doi:https://doi.org/10.1016/j.is.2014.07.006

Hilbert, M. (2016). Big data for development: a review of promises and challenges. *Development Policy Review*(34).

Jensen, M. (2013). *Challenges of Privacy Protection in Big Data Analytics*. Paper presented at the 2013 IEEE International Congress on Big Data.

Johnsen, F. (2015). Hva eller hvem er Hadoop? Retrieved from http://www.cw.no/artikkel/big-data/hva-hvem-hadoop

Kacfah Emani, C., Cullot, N., & Nicolle, C. (2015). Understandable Big Data: A survey. *Computer Science Review, 17*, 70-81. doi:10.1016/j.cosrev.2015.05.002

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). *Big Data: Issues and Challenges Moving Forward*. Paper presented at the 2013 46th Hawaii International Conference on System Sciences.

Kommunal- og moderniseringsdepartementet. (2017). Stordata og offentlige tjenester. Retrieved from https://www.regjeringen.no/contentassets/57d2bce6c0b3481ab92659bf015cad71/prg_bakgrunnsinfo.pdf

Kshetri, N. (2014). Big data′s impact on privacy, security and consumer welfare. *Telecommunications Policy, 38*(11), 1134-1145. doi:10.1016/j.telpol.2014.10.002

Leavitt, N. (2013). Storage Challenge: Where Will All That Big Data Go? *Computer, 49*(9), 4. doi:10.1109/MC.2013.326

Lei, X., Chunxiao, J., Jian, W., Jian, Y., & Yong, R. (2014). Information Security in Big Data: Privacy and Data Mining. *IEEE Access, 2*, 1149-1176. doi:10.1109/access.2014.2362522

Liu, S. M., & Yuan, Q. (2015). The Evolution of Information and Communication Technology in Public Administration. *Public Administration and Development, 35*(2), 140-151. doi:10.1002/pad.1717

Lord, N. (2017). What is GDPR? Understanding and complying with GDPR data protection requirements. . Retrieved from https://digitalguardian.com/blog/what-gdpr-general-data-protection-regulation-understanding-and-complying-gdpr-data-protection

Maciejewski, M. (2016). To do more, better, faster and more cheaply: using big data in public administration. *International Review of Administrative Sciences, 83*(1_suppl), 120-135. doi:10.1177/0020852316640058

Mai, J.-E. (2016). Three models of privacy. *Nordicom Review, 171-175*(37), 175. doi:10.1515/nor-2016-0031

Mantelero, A., & Vaciago, G. (2015). Data protection in a big data society. Ideas for a future regulation. *Digital Investigation, 15*, 104-109. doi:10.1016/j.diin.2015.09.006

McAfee, A., & Brynjolfsson, E. (2011). Big Data the Management Revolution. *Harvard Business Review*, 10.

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of Big Data Privacy. *IEEE Access, 4*, 1821-1834. doi:10.1109/access.2016.2558446
Merriam-Webster. (2018). Anonymize. Retrieved from https://www.merriam-webster.com/dictionary/anonymize
Merriam-Webster. (2018). Profiling. Retrieved from https://www.merriam-webster.com/dictionary/profiling
Miller, G., & Mork, P. (2013). From Data to decisions a value chain for big data. *IEE Computer Society*, 59.
Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big Data: Prospects and Challenges. *Vikalpa, 40*(1), 74-96. doi:10.1177/0256090915575450
Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., . . . Kaye, J. (2018). Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law & Security Review, 34*(2), 222-233. doi:10.1016/j.clsr.2018.01.002
NAV. (2018). Kva er NAV? Retrieved from https://www.nav.no/no/NAV+og+samfunn/Om+NAV/Fakta+om+NAV/kva-er-nav
NAV. (2018). Organisering av NAV. Retrieved from https://www.nav.no/no/NAV+og+samfunn/Om+NAV/Fakta+om+NAV/organisering-av-nav
Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems, 24*(1), 3-14. doi:10.1016/j.jsis.2015.02.001
Oates, B. J. (2006). *Researching information systems and computing*. London ; Thousand Oaks, Calif.: SAGE Publications.
Oxforddictionaries. (2018). Bias. Retrieved from: https://en.oxforddictionaries.com/definition/bias
Perera, C., Ranjan, R., & Lizhe, W. (2015). End-to-End Privacy for Open Big Data Markets. *The IEE computer society*, 10.
Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314-347. doi:10.1016/j.ins.2014.01.015
Pouloudi, L. D. I. a. A. (1999). Privacy in the information age stakeholders, interests and values. *Journal of Business Ethics, 27-38*(22), 38.
Reusch, M. (2017). hjemmel. Retrieved from https://snl.no/hjemmel
Rogge, N., Agasisti, T., & Witte, K. D. (2017). Big data and the measurement of public organizations' performance and efficiency: The state-of-the-art. *Public Policy and Administration, 32*(4), 19. doi:10.1177/0952076716687355
Rouse, M. (2015). Data Warehouse. Retrieved from https://searchsqlserver.techtarget.com/definition/data-warehouse
Rouse, M. (2016). big data analytics. Retrieved from https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Sagioglu, S., & Sinanc, D. (2013). Big Data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 6. doi:10.1109/CTS.2013.6567202

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Mol Syst Biol, 8*, 612. doi:10.1038/msb.2012.47

Skatteetaten. (2018). Samfunnsoppdrag og strategi. Retrieved from https://www.skatteetaten.no/om-skatteetaten/om-oss/samfunnsoppdrag-strategi/

Soria-Comas, J., & Domingo-Ferrer, J. (2015). Big Data Privacy: Challenges to Privacy Principles and Models. *Data Science and Engineering, 1*(1), 21-28. doi:10.1007/s41019-015-0001-x

Tan, Q., & Pivot, F. (2015). *Big Data Privacy: Changing Perception of Privacy*. Paper presented at the 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity).

Taneja, H., Kapil, & Singh, A. K. (2015). Preserving Privacy of Patients Based on Re-identification Risk. *Procedia Computer Science, 70*, 448-454. doi:10.1016/j.procs.2015.10.073

Tankard, C. (2012). Big data security. *Network Security, 2012*(7), 5-8. doi:10.1016/s1353-4858(12)70063-6

Techopedia. (2018). Web Scraping. Retrieved from: https://www.techopedia.com/definition/5212/web-scraping

Techopedia. (2018). Data De-identification. Retrieved from: https://www.techopedia.com/definition/25010/data-de-identification

Tomter, L., & Remen, A. C. (2017). Datatilsynet fant lovbrudd: Millionbøter etter outsourcing av sykehus-IT. *NRK*. Retrieved from https://www.nrk.no/norge/millionboter-etter-outsourcing-av-sykehus-it-1.13751516

van Loenen, B., Kulk, S., & Ploeger, H. (2016). Data protection legislation: A very hungry caterpillar. *Government Information Quarterly, 33*(2), 338-345. doi:10.1016/j.giq.2016.04.002

Victor, N., Lopez, D., & Abawajy, J. H. (2016). Privacy models for big data: a survey. *International Journal of Big Data Intelligence, 3*(1). doi:10.1504/ijbdi.2016.073904

Yin, R. K. (2009). *Case study research : design and methods* (4th ed.). Los Angeles, Calif.: Sage Publications.

Yu, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access, 4*, 2751-2763. doi:10.1109/access.2016.2577036

Zeng, W., Yang, Y., & Luo, b. (2013). *Access control for big data using data content*. Paper presented at the IEEE International Conference on Big Data.

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems-the International Journal of Escience, 28*(3), 10. doi:doi:10.1016/j.future.2010.12.006

9. Vedlegg

Vedlegg A - Intervjuguide NAV, Skatteetaten og SSB

Spørsmål relatert til bruken av stordata i organisasjonen

1. Hvordan jobber du med stordata? Hva er dine arbeidsoppgaver?
2. Hvordan jobber organisasjonen med stordata?
3. Hvilke prosjekter jobber dere med innenfor stordata.
4. Hvor lenge har organisasjonen holdt på med stordata-prosjekter?
5. Hvordan har prosjektene innenfor stordata gått? Hva har dere lært?

Spørsmål relatert til personvern og stordata

1. Hvilke tiltak gjør for å opprettholde personvern i stordata?
2. Hva anser dere som de største utfordringene med tanke på personvern og stordata?

Innsamling av data

1. Hvordan samles stordata i organisasjonen? Fra hvilke kilder.
2. Hvor store mengder med data samles inn?
3. Er brukere informert at det samles inn data om dem? Hvordan informeres dem? Mtp. Samtykke.
4. Hvor mye data er tilstrekkelig å samle inn om personer for at det kan utnyttes i forhold til at dere kan få utbytte av det?
5. Blir mye data som samles inn ikke utnyttet? Er det planer for disse? Samles data inn for senere bruk?
6. Hva ser dere på som sensitive data?
7. Hvordan er retningslinjene for å publisere open data?
8. Kan datakvaliteten ha noe å si for personvern?

Analyse av data

1. Hvordan anvendes dataene? Bruker dere programvare? Analyser?
2. Hvilke personvernutfordringer ser dere er sentrale ved analyse/prosessering av stordata?
3. At enkeltpersoner kan identifiseres er sett på som en av de største utfordringene ved stordata. Har dere tiltak for å hindre de-anonymisering av data?
4. Bruker organisasjonen eksterne leverandører for stordata-løsninger?
5. Hvilke teknikker bruker dere for anonymisering?
6. Hvilke personvern-utfordringer kommer i utvikling av stordata-modeller?
7. Hvilke personvernutfordringer er sentrale i din rolle/din stilling?

Tilgang og ansvarsområder

1. Hvem har ansvaret for dataen i organisasjonen i sin helhet? Hvem har tilgang til dataene, utviklere, konsulenter, prosjektledere o.l.?
2. Brukes dataene i mellom ulike avdelinger i organisasjonen?
3. Er det andre organisasjoner som tar i bruk dataene? Sendes de til andre tjenester?
4. Bruker dere data fra andre organisasjoner/bedrifter/tjenester?

Lagring av stordata

1. Hvor lagres dataene? Hvor vil de lagres i fremtiden?
2. Hvordan lagres dataene? Er de kryptert?
3. Hvilke metoder brukes for å sikre dataene som blir lagret?
4. Bruker organisasjonen leverandører for lagring? F.eks skytjenester.

GDPR

1. Hvordan jobber dere med GDPR?
2. Hvilke tanker har dere gjort rundt GDPR?
3. Vil stordata brukes til å kunne utarbeide profiler om meg som enkeltperson?
4. Har man rett til å sette seg imot profilering?
5. Er folk klar over hva som samles om de per tidspunkt?

Lover og regler for personvern

1. Har dere noen utfordringer med tanke på lover og regler?
2. Hvilke tiltak har dere gjort med tanke på GDPR?
3. Hva er de største utfordringene med tanke på GDPR?
4. Vil organisasjonen greie å fylle alle kravene før GDPR trer i kraft?

Vedlegg B - Intervjuguide Datatilsynet

Spørsmål relatert til bruken av stordata i organisasjonen

1. Hva er dine arbeidsoppgaver? Hvordan jobber du med personvern/stordata?
2. Hva er Datatilsynets rolle når det gjelder personvern og stordata?
3. Hvilke offentlige organisasjoner jobber dere med mtp. stordata og personvern?

Spørsmål relatert til personvern og stordata

1. Hvilke tiltak bør organisasjoner gjøre for å opprettholde personvern i stordata?
2. Hva anser dere som de største utfordringene med tanke på personvern og stordata?
3. Går dere inn hos organisasjoner for å kontrollere om de følger lover/regler?
4. Hva er konsekvensen for å bryte lover/regler?
5. Har det hendt at dere har måttet sette foten ned når det gjelder stordata prosjekter pga . utfordringer mtp personvern? Hva var grunnen?

Innsamling av data

1. Er det noen begrensninger på hvordan data kan samles inn om brukere? Hva tenker dere om f.eks data som samles inn fra sosiale medier?
2. Tenker dere at mengden med data som samles inn kan ha noe å si for personvernet?
3. Hvordan tenker Datatilsynet at brukere bør informeres at data samles inn om dem? Mtp. Samtykke.
4. Hva ser dere på som sensitive data?
5. Kan datakvaliteten/bias ha noe å si for personvern?
6. Opplever dere at offentlige organisasjoner følger reglene for innsamling av data?

Analyse av data

1. Hvilke metoder for analyse av stordata brukes for å sikre personvern?
2. Ser dere utfordringer ved å benytte tredjepartsløsninger eller organisasjoner for å analysere data?
3. Hvilke personvernutfordringer ser dere er sentrale ved analyse/prosessering av stordata?
4. Hvilke tanker har dere om at individer kan bli identifisert? Hva er tiltak dere mener kan gjøres for å hindre det?
5. Hvilke personvernutfordringer er sentrale i din rolle/din stilling?
6. Hva tenker dere om publisering av big open data? Hvor ligger grensen for hva som kan publiseres? Hva blir for sensitivt?

Tilgang og ansvarsområder

1. Hva tenker dere om tilgangskontroll? Skal en regnskapsfører ha tilgang til stordata som ikke er relevant for han/hun?
2. Jobber dere med kontroll av tilgangskontroll?

Lagring av stordata

1. Hva tenker datatilsynet er gode retningslinjer for lagring av stordata? Spesielt mtp. sensitive data.
2. Hvilke krav stiller dere til lagring og kryptering av stordata?
3. Hva tenker dere er utfordringer ved å benytte tredjeparter til lagring av data? F.eks skytjenester.

GDPR

1. Hvordan jobber dere med GDPR?
2. Hvilke tanker har dere gjort rundt GDPR?
3. Har dere kontroll på organisasjoners fremgang mtp. GDPR?
4. Hva er datatilsynets rolle mtp. GDPR?
5. Hva er de største utfordringene mtp. GDPR?
6. Hva tenker dere om organisasjoner som ikke vil klare å oppfylle kravene før lovverket trer i kraft?

Vedlegg C - Forstudie

"Hvordan håndterer norske offentlige organisasjoner personvern når de anvender stordata?"

Bakgrunnsinformasjon om ansatte vi intervjuer

- Hvilken organisasjon de er ansatt hos.
- Stilling til personen.
- Personens rolle i organisasjonen, hva de arbeider med.

Hoveddel

Spørsmål og temaer rettet til ledelsen av organisasjonen

1. Hva organisasjonen jobber med, deres samfunnsoppgave.
2. Hvordan organisasjonen jobber med stordata og personvern.
3. Hvor lenge har organisasjonen holdt på med stordata prosjekter.
4. Hvilke prosjekter jobber dere med innenfor stordata.
5. Hvilke tiltak organisasjonen gjør for å opprettholde personvern i stordata.
6. Hvordan har prosjektene innenfor stordata gått? Hva har dere lært?
7. Hvordan bruker organisasjonen stordata for å utvikle tjenestene sine?
8. Hvilke tiltak de planlegger i fremtiden for å opprettholde personvern i stordata .
9. Hva anser de som de største utfordringene med tanke på personvern og stordata .
10. Hvordan har organisasjonen kommet frem til at tiltakene om personvern er de beste/ at de fungerer?

Lover og regler for personvern

11. Hvordan de forholder seg til lover og regler.
12. Hvilke tiltak har de gjort med tanke på GDPR.
13. Hva er de største utfordringene med tanke på GDPR?
14. Vil organisasjonen greie å fylle alle kravene før GDPR trer i kraft.

Ansattes arbeidsoppgaver i henhold til stordata

15. Hva er dine arbeidsoppgaver i forhold til personvern og stordata.
16. Spørsmål om hva personen jobber med i relasjon til stordata.
17. Hva personen gjør for å opprettholde personvern i henhold til sine arbeidsoppgaver

Innsamling av stordata

18. Hvordan samles stordata i organisasjonen? Fra hvilke kilder.
19. Blir mye data som samles inn ikke utnyttet? Er det planer for disse? Samles data inn for senere bruk?
20. Hvor mye data er tilstrekkelig å samle inn om personer for at det kan utnyttes i forhold til trade-off?

21. Hvor store mengder med data samles inn?

Lagring av stordata

22. Hvor lagres dataene? Hvor vil de lagres i fremtiden?
23. Hvordan lagres dataene? Er de kryptert? Hvordan?
24. Hvilke metoder brukes for å sikre dataene som blir lagret?
25. Hvilke data ut i fra de kildene blir utnyttet/lagret?
26. Bruker organisasjonen leverandører for lagring? F.eks skytjenester.

Anvendelse av stordata

27. Hvordan anvendes dataene? Bruker dere programvare? Analyser?
28. Har dere tiltak for å hindre de-anonymisering av data?
29. Bruker organisasjonen eksterne leverandører for stordata-løsninger?

Tilgang og ansvarsområde

30. Hvem har ansvaret for dataen i organisasjonen i sin helhet? Hvem har tilgang til dataene, utviklere, konsulenter, prosjektledere o.l.?
31. Brukes dataene i mellom ulike avdelinger i organisasjonen?
32. Er det andre organisasjoner som tar i bruk dataene? Sendes de til andre tjenester?
33. Bruker dere data fra andre organisasjoner/bedrifter/tjenester?