SINTEF

SINTEF A25786 - Åpen

Rapport

Nye datakilder i evaluering av store statlige investeringer – et potensial for Big Data?

Forfattere
Heidi Bull-Berg og Nils Olsson

SINTEF Teknologi og samfunn
Anvendt økonomi
2014-01-06

SINTEF

SINTEF Teknologi og samfunn

Postadresse:
Postboks 4760 Sluppen
7465 Trondheim

Sentralbord: 73593000
Telefaks: 73590260

ts@sintef.no
www.sintef.no
Foretaksregister:
NO 948007029 MVA

Rapport

Nye datakilder i evaluering av store statlige investeringer – et potensial for Big Data?

EMNEORD:
Big Data
Evaluering
Statlige investeringer

VERSJON
1

DATO
2014-01-06

FORFATTER(E)
Heidi Bull-Berg, Nils Olsson

OPPDRAGSGIVER(E)
Concept-programmet

OPPDRAGSGIVERS REF.
Gro Holst Volden

PROSJEKTNR
102003719

ANTALL SIDER OG VEDLEGG:
41 + vedlegg

SAMMENDRAG

Overskrift sammendrag

Denne rapporten kartlegger og drøfter hvordan Big Data kan brukes til evaluering av store statlige investeringer. I vårt tilfelle med fokus på evaluering er vi egentlig interessert i «nye data», ikke bare de ekstremt store. Vi finner få publikasjoner som eksplisitt omtaler Big Data relatert til evaluering, men det finnes flere eksempler på analyser som hadde kunnet vært brukt i evalueringssammenheng. Big Data kan deles inn i internettaktivitet, bevegelsesrelaterte data, data om fysiske omgivelser og kommersiell aktivitet. I tillegg finnes ulike former for interne data i flere etater som har potensiale til å brukes i evalueringer. Rapporten diskuterer spesielt følgende aspekter som det må tas hensyn til ved bruk av Big Data i evaluering; personvern, tilgjengelighet, anvendbarhet og relevans, eiendomsrett, kostnader og kompetanse. Vi ser at nye og større datamengder bør være interessante til bruk i både tidligfase av store statlige investeringer som beslutningsgrunnlag (ex ante), og i evaluering av ferdigstilte tiltak (ex post).

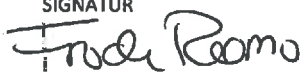
UTARBEIDET AV
Heidi Bull-Berg

SIGNATUR


KONTROLLERT AV
Håkon Finne

SIGNATUR


GODKJENT AV
Frode Rømo

SIGNATUR


RAPPORTNR
SINTEF A25786

ISBN
978-82-14-05660-0

GRADERING
Åpen

GRADERING DENNE SIDE
Åpen

Sammendrag

Målet for denne studien har vært å kartlegge og drøfte hvordan Big Data kan brukes til evaluering av store statlige investeringer. Forskningsprogrammet Concept, som finansierer denne studien, driver følgeforskning på store statlige investeringsprosjekter. Forskerne er blant annet opptatt av hvordan prosjektene kan etterevalueres, med fokus på realisering av bruker- og samfunnseffekter. En av erfaringene så langt er at det kan være krevende å få tilgang til data som direkte måler det en er ute etter. I evaluering er man dessuten opptatt av hvordan ulike typer data og analyse kan kombineres (triangulering).

Vi diskuterer den nye og voksende trenden Big Data, som kjennetegnes av store mengder data som genereres gjennom bruk av nye teknologiske løsninger og systemer. Vi diskuterer videre bruk av Big Data i evaluering, med spesielt fokus på hvordan denne typen data kan benyttes i evaluering av store statlige investeringer. Vi ser at nye og større datamengder kan være interessant til bruk i både tidligfase av store statlige investeringer som beslutningsgrunnlag (ex ante), og i evaluering av ferdigstilte tiltak (ex post). I vårt tilfelle med fokus på evaluering er vi egentlig interessert i «nye data», ikke bare de ekstremt store. Behovet for å avgrense studien har gjort at vi i praksis har fokusert mest på vei-, jernbane- og byggeprosjekter, men mye av det som står i rapporten gjelder generelt.

Resultatene som presenteres i rapporten er fremkommet gjennom litteraturstudier og samtaler med det vi har ansett som relevante personer i Norge når det gjelder Big Data generelt, og mer spesifikt til bruk i prosjektevaluering.

Big Data er datasett som er så store at de ikke er egnet til å hverken innhente, lagre, prosessere eller analysere ved hjelp av tradisjonelle databaseverktøy. Tradisjonelle kjennetegn er volum, hastighet, variasjon og aktualitet. Bruk av Big Data handler om å hente ut innsikt for å kunne ta kunnskapsbaserte avgjørelser. Det store fortrinnet er muligheten til å koble utallige datakilder for å se nye sammenhenger, mønstre, effekter mv. Det er mulig å "oppdage ting man ikke visste man lette etter". Man kan prosessere nye typer data og utnytte ustrukturert informasjon.

Det har vært en rask utvikling innenfor området Big Data de senere år. Følgende viktige utviklingstrekk er verdt å merke seg:

- Større mengder data, inkludert data fra internett og utviklingen av sensor- og sporingsteknologi
- Økt tilgjengelighet
- Økt press for å gjøre data mer tilgjengelig
- Tilgang til lagrings- og analysekapasitet gjennom nye løsninger for systemarkitektur og skytjenester med stor og skalerbar kapasitet til en lav kostnad.
- Tilgang til IT-plattformer for å sette data inn i en sammenheng, eksempelvis digitale kart for presentasjon av posisjonsdata, eller bygningsinformasjonsmodeller (BIM)

Big Data kan deles inn i følgende kategorier etter måten de samles inn/genereres på:

- Internettaktivitet, inkludert aktivitet på sosiale media og data fra søkemotorer (cookies, tekst, klikk)

- Bevegelsesrelaterte data, inkludert GPS, mobiltrafikk og lokalisering, bomstasjoner, RFID-brikker på gods
- Data om fysiske omgivelser
- Kommersiell aktivitet og bruk av betalingstjenester

I tillegg merker vi at det finnes ulike former for interne data i flere etater som har potensiale til å brukes mer aktivt enn hva som har vært tilfelle hittil ved evalueringer. Disse data finnes blant annet i systemer for vedlikeholdsoppfølging. Omfanget av slike data er foreløpig ikke så store at de dekkes under Big Data, men bruken av slike data synes å ha mye fellestrekk med det som brukes i Big Data.

Bruk av Big Data synes å ha størst utbredelse innenfor analyser basert på internettaktivitet. Denne type analyse synes ikke å være den mest aktuelle til evaluering av store statlige investeringer, men det kan være aktuelt å studere hvordan eksempelvis det nye Operabygget blir omtalt på internett gjennom sentiment-analyse.

Bevegelsesrelaterte data synes meget aktuelle til evaluering av store statlige investeringer, både innenfor samferdsel og bygninger. Ved bruk av slike data kan bevegelsesmønster for brukere (og for så vidt ikke-brukere) av nye samferdselsinvesteringer og nye bygninger kartlegges. Grunnet personvernhensyn må analysene sannsynligvis gjøres for større grupper av brukere.

Ulike sensordata kan være aktuelle, spesielt til kvalitetssikring og komplettering av bevegelsesdata, og som mulige forklaringsfaktorer. Dette kan være temperatur eller energibruk i bygninger eller værdata og hastighetslogger fra togmateriell for samferdselstiltak.

Norge har gode statistiske registre for flere områder som befolkning, sysselsetting, bosetting mv. Disse kan defineres som Big Data, men vil av noen utelukkes som dette da de er utarbeidet kun til statistiske formål. Registrene er likevel svært aktuelle både i seg selv og når det gjelder å koble sammen ulike datakilder, som jo er selve nøkkelen når det gjelder Big Data.

Vi finner få publikasjoner som eksplisitt omtaler Big Data relatert til evaluering. Det er allikevel publisert flere eksempler på analyser som kunne blitt brukt inn i en evaluering. To eksempler er bruk av mobiltelefonregistreringer for å beskrive transportmønster, og bruk av logging av smarttelefoner til å vise bevegelsesmønster i museum. Data fra mobiltelefoner ble registrert fra en million brukere i Boston i løpet av tre måneder. Disse data ble koblet mot demografiske data og benyttet til å kartlegge faktisk transportbehov. Både mobildata og demografiske data var aggregerte for å ivareta personvernhensyn. I det andre eksempelet ble enheter med blåtann-sender logget, i praksis smarttelefoner, og brukt til å kartlegge besøkene i Louvre-museet. Dermed kunne man beskrive blant annet besøkernes bevegelsesmønster og lengde på besøkene. Teknologien som er brukt i disse eksemplene hadde vært aktuell å bruke ved evaluering av samferdselsprosjekter og offentlige (og andre) bygninger.

Rapporten diskuterer videre utvalgte utfordringer relatert til bruk av Big Data i evaluering som vi mener er viktige; personvern, tilgjengelighet, anvendbarhet og relevans, eiendomsrett, kostnader og kompetanse.

Tilgjengelighet

Tilgjengeligheten av data styres av to forhold. For det første må noen etterspørre data. Det finnes flere eksempler på at data har vært tilgjengelige, men ikke blitt brukt fordi ingen så potensialet til anvendelse, og derfor ikke etterspurte disse data. Det andre forholdet rundt tilgjengelighet er om data kan utleveres. Det er en pågående trend for offentliggjøring av data. Det synes sannsynlig at Concept-programmet og andre som arbeider med evaluering av store statlige investeringer kan påvirke tilgjengeligheten av data ved å etterspørre data og arbeide for tilgjengeliggjøring av data som er interessante til evaluering. Det synes å være uavklart om offentlighetsprinsippet i offentlig sektor gjelder for data.

En utfordring i mange evalueringssituasjoner er å få data som dekker lange tidsperioder, og spesielt data som beskriver situasjonen før et prosjekt starter. Dette kan være mange år tilbake i tiden når en ex-post evaluering skal utføres. Det kan bli nødvendig å iverksette tiltak for å sikre at data blir lagret over lange tidsrom, slik at data er tilgjengelige i en evalueringssituasjon.

Anvendbarhet

Big Data skaper nye muligheter til å analysere et fenomen basert på ulike typer av data. Dette øker mulighetene for en evaluator til å finne indikatorer som er relevante i forhold til det tiltaket man evaluerer.

Big Data (eller store data) er anvendbare på flere måter. Spesielt kan flere ulike datasett som belyser samme fenomen brukes til triangulering og kvalitetssikring av evalueringer. Trianguleringen kan inkludere bruk av etablerte typer av informasjon, som intervjuer og dokumentanalyse. Big Data vil trolig også kunne komplettere og forbedre eksisterende evalueringsparametere, samt bidra med nye parametere for å synliggjøre virkninger som tidligere ikke har latt seg måle.

Det finnes flere eksempler der sensordata fra ulike systemer og ulike målingsprinsipper kan brukes for å belyse samme fenomen. Triangulering og kvalitetssikring av data og analyser kan også gjøres basert på helt ulike typer av data.

Big Data har også potensiale til å gi ny informasjon om folks betalingsvillighet for ulike tiltak. Dette er et sentralt element i samfunnsøkonomiske analyser. Studier av betalingsvillighet innfor transport har de seneste 10-årene til stor del vært basert på spørsmål til brukere (SP-studier). Big Data åpner for nye typer studier basert på individers faktiske valg. I forlengelsen kan Big Data på denne måten bidra til å prissette betalingsvillighet innen sektorer det det hittil har vært utfordrende å prissette samfunnsnytten av store statlige investeringer.

Relevans

Big Data er gjerne samlet inn på en utradisjonell måte for en statistiker. Dette medfører at man trenger nye statistiske metoder for å forstå data som ikke er perfekte, og som ikke er samlet inn til et statistisk formål, men som likevel har et potensiale til å brukes. Man må tenke nytt når det gjelder bruk av vitenskapelige metoder. Tradisjonelle statistiske problemstillinger som representativitet, signifikans, utvalgskriterier, frafall etc. må tilpasses de nye typene av data. Vi tror at disse begrep fortsatt vil være relevante, men tror at endring i vitenskapelige metoder vil bli mer aktuelt når Big Data-analyser blir mer etablerte. Et spørsmål er om bruk av Big Data i fremtidens beslutningsprosesser i det hele tatt kan sammenliknes med bruk av mer tradisjonelle statistiske kilder.

En annen utfordring knyttet til bruk av Big Data i evaluering er at sammenliknbarhet over tid kan være vanskelig. Dette gjelder spesielt data basert på internettbruk, da det har vært store endringer i internettbruken de seneste årene. Endring i bruk henger sammen med at brukermassen også har endret seg. Facebook ble for eksempel i sin tidlige fase i hovedsak benyttet av ungdom, mens vi ser at mediet i dag benyttes av både gamle og unge. Sensordata og data fra kommersielle transaksjoner synes å være mer sammenlignbare over tid. Posisjonsdata kan være påvirket av hvilken teknologi som brukes for å registrere posisjon og bevegelse. Utfasing ev en teknologiplattform (som det gamle mobiltelefonsystemet NMT900) og innfasing av en annen (som smarttelefoner, eller telefoner med blåtann) kan skape utfordringer ved sammenligning av data over lange tidsperioder. Dette innebærer at analyser basert på Big Data kan være mer relevante til å beskrive situasjonen ved tidspunktet for evaluering, sammenlignet med å beskrive utviklingen over en lengre periode. Men når det er sagt så har andre datakilder for evaluering, som intervjuer og bruk av veletablerte målinger, også svakheter når det gjelder å nøyaktig og objektivt beskrive forskjellen mellom situasjonen for og etter et større tiltak. Disse utfordringene kan reduseres dersom data lagres med høyest mulig oppløsning, og det angis tydelig hvordan dataene er innsamlet og bearbeidet.

Personvern

Personvern trenger ikke å være til hinder for bruk av Big Data, selv om det synes å være det tema som folk flest først tar opp i tilknytning til bruk av Big Data. Når det gjelder innsamling og analyse av Big Data med personvernopplysninger står to rettslige grunnlag sentralt i personopplysningsloven: lovhjemmel og krav om samtykke. Alle data som ikke inkluderer personopplysninger er i utgangspunktet uproblematisk, både som enkeltstående datakilder og til kombinasjon av flere kilder. Data fra ulike kilder kan kombineres uten at personopplysninger nødvendigvis blir avslørt, men dette kan være utfordrende hvis man jobber med mange ulike detaljerte datasett. Anonymisering av data kan gjøres på høyoppløselige data, eller ved aggregeringer av data, der hver gruppe inkluderer så mange individer at enkeltindivider ikke kan identifiseres.

Data skal i utgangspunktet benyttes til det formålet de var tiltenkt (formulert på forhånd). Vårt inntrykk er at når data blir anonymiserte (for eksempel aggregerte) gjelder ikke disse reglene. Aggregerte data er i utgangspunktet ikke et problem i forskningssammenheng, når man ønsker å avdekke trender, mønstre etc. Det samme gjelder i evalueringssammenheng.

Personvernproblematikken synes håndterbar, men forutsetter tilgang til teknisk og juridisk kompetanse, som kan medføre ekstra kostnader, og legge begrensninger på hvilken oppløsning eller detaljeringsgrad som analysene kan utføres på.

Eiendomsrett

Med eiendomsrett tenker vi her på hvem som faktisk sitter på rettighetene til informasjonen som samles inn, være seg med eller uten personopplysninger. Lovverket for eiendomsrett til Big Data synes ikke å være avklart. To prinsipper som flere nevner er at (1) den som samlet inn dataene eier dem og (2) aggregerte data eies av den som har utført aggregeringen. Et annet viktig tema i dag og spesielt fremover vil være sporbarhet i bruken av data. Hvem har brukt de, hvem har hatt innsyn? Det vil i fremtiden bli viktig å kunne ha gode og pålitelige systemer rundt dette. Utfordringene vedr. eiendomsrett (og øvrige juridiske forhold) øker dersom data hentes eller overføres til andre land. Dette er en aktuell problemstilling når mye av

analysene utføres som sky-baserte tjenester. Det er usikkert hvorvidt eiendomsrett vil være en barriere for bruk av Big Data i evaluering. Tilsvarende som for personvern trengs juridisk kompetanse for å avklare de ulike forhold.

Kostnad

Kostnader for bruk av Big Data er redusert da både lagrings- og analysekapasitet har blitt billigere og lettere tilgjengelig ved bruk av Hadoop[1]-teknologi og skybaserte løsninger. Også innsamling av data har blitt billigere enn før. Sensorer er lettere tilgjengelig, er billig hyllevare og enklere i installasjon og drift. Dersom bruk av Big Data supplerer tidligere kilder vil de samlede kostnadene trolig øke. Nytten av å få disse data vil likevel kunne være stor nok til å forsvare kostnadene. I det tilfelle at Big Data kan erstatte mer tid- og kostnadskrevende innhenting av data til evaluering så representerer det en effektivitetsøkning. Man kan derved gjøre mer evaluering for de samme pengene, eksempelvis evaluere flere tiltak, eller redusere kostnaden for evaluering. Det er en forventning om at data kommer til å få en kommersiell verdi, men også at volumet av åpne data kommer til å øke. Mye av de data som er aktuelle til evaluering av store statlige investeringer bør være i den åpent tilgjengelige gruppen. Dette vil kunne senke kostnadene ved datainnhenting. Kompetanse for analyse og prosessering av data er derimot nødvendig, og kan være en kostnadskomponent i seg selv.

Kompetanse

Tilgang til kvalifiserte folk med riktig kompetanse innen blant annet maskinlæring oppgis å være en flaskehals for utviklingen innen området. I tillegg til IT-kompetanse trengs også svært god kunnskap på analyse og visualisering av data, samt forskningskompetanse, både for å forstå hvordan man rent teknisk kan prosessere data, men også hvilken kunnskap man kan hente ut av dem.

Våre intervjuobjekter mener at Norge ikke er langt fremme på analyse og prosessering av data, selv om flere IT-miljøer har begynt å bygge kompetanse på området den siste tiden, som for eksempel Telenor. Manglende kompetanse vil derfor også være en barriere for utnyttelsen av Big Data i Norge. For å løse dette er det viktig at både offentlig sektor og privat næringsliv er klar over denne utfordringen, og er i forkant av behovene når de oppstår både ved å tilrettelegge for utdanning og gjennom målrettet rekruttering.

Det synes som at det finnes store muligheter for bruk av nye (store) data i evaluering. Concept kan bli blant de første til å vise muligheter i Big Data i forhold til evaluering generelt, og spesielt relatert til store statlige investeringer. Vi har funnet flere eksempler på kreativ bruk av Big Data som er relevant for evaluering av store statlige investeringer. Når det gjelder de to spesifikke sektorområdene samferdsel og bygg konkluderer vi med at det her finnes muligheter for å fremskaffe et bredere datagrunnlag (både basert på Big Data og andre nye datakilder), som trolig vil kunne både forbedre tidligere måleparametere, samt tilføre nye i evaluering av tiltak. En skjematisk vurdering av aktuell kilder er gitt i vedlegg C. Tabellen er på

[1] Hadoop er et gratis, Java-basert programmeringsrammeverk som støtter prosessering av store datamengder i et distribuert databehandlings miljø. Hadoop kjører dataintensive applikasjoner gjennom MapReduce parallell prosessering.

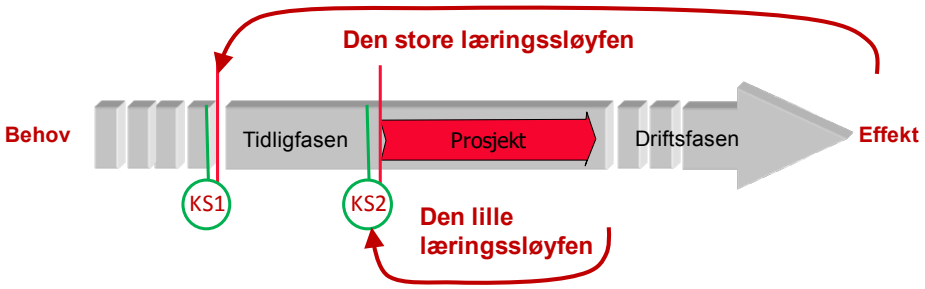
ingen måte uttømmende, men kan betraktes som et utgangspunkt for videre arbeid med bruk av Big Data i evaluering av tiltak innenfor samferdsel og bygg.

Vi anbefaler at Concept bidrar til å starte opp piloter som ser på mulighetene for utnyttelse av Big Data i evaluering av store statlige investeringstiltak innenfor samferdsel og bygg.

Innholdsfortegnelse

1 Innledning og formål med studien

Finansdepartementets kvalitetssikringsordning (KS) gjelder for store statlige investeringstiltak over 750 mill. kroner (Concept 2013). Hovedtyngden av de investeringer som er underlagt KS er investeringer innenfor sektorene forsvar, samferdsel og bygninger, men også større IT-tiltak. Ordningen har i dag 2 formelle beslutningspunkter, KS1 (før vedtak i regjering) og KS2 (før vedtak i storting) som vist i Figur 1. KS1 handler om tidligfasevurderinger (blant annet samfunnsøkonomisk analyse, konseptvalg), KS2 handler om prosjektfasen (blant annet kostnadsestimering og risikovurdering). Concept-programmet driver følgeforskning på prosjektene som er en del av ordningen, samt på selve ordningen (metodeutvikling, effekten av den etc.). Det antas å være hensiktsmessig med ulike former for evaluering av de statlige investeringene (ex post og ex ante). Dette vil styrke den store læringssløyfen som vist i Figur 1.



Figur 1. Kvalitetssikringsregimet (Concept 2013)

Direktoratet for økonomistyring definerer evaluering som "en systematisk datainnsamling, analyse og vurdering av en planlagt, pågående eller avsluttet aktivitet, en virksomhet, et virkemiddel eller en sektor". Evalueringer kan gjennomføres før et tiltak iverksettes (ex ante), underveis i gjennomføringen, eller etter at tiltaket er avsluttet (ex post). Evalueringen kan utføres av interne eller eksterne fagmiljøer. Generelt kan man også si om evalueringer at de:

- Vurderer verdien av noe, i tillegg til å beskrive
- Stiller krav om systematikk og metode som det må kunne redegjøres for
- Innebærer at noen betrakter og "stiller seg utenfor" det som skal evalueres
- Forutsetter at det finnes et faktagrunnlag

Data til bruk i evaluering kan eksempelvis deles i følgende kategorier:

- Data som finnes og kan innhentes med eksisterende datakilder
- Data som finnes, men som av ulike grunner ikke gjøres tilgjengelig
- Data som i dag ikke finnes men som bør kunne genereres med kjent teknologi

Tilgang på gode relevante data kan være en utfordring ved evaluering av store statlige investeringer (Volden og Samset 2013). Dette kan synes som et paradoks, når omfanget av data generelt sett øker. Dette

har vært utgangspunktet for å se på potensialet for å bruke Big Data i evaluering av store statlige investeringer i denne studien.

Utgangspunktet for denne studien er å se nærmere på denne nye og voksende trenden; store mengder data som genereres gjennom bruk av nye teknologiske løsninger og systemer. Målet er å kartlegge og drøfte i hvilken grad denne typen data kan benyttes i evaluering av store statlige investeringstiltak. Med statlig investeringsvirksomhet menes de investeringer som finansieres, planlegges og gjennomføres av sentralforvaltningen. Det er imidlertid en omfattende interaksjon mellom lokal, regional og sentral forvaltning i forarbeidene til statlige investeringstiltak og i planleggingen av slike tiltak. Et eksempel er interaksjon rundt reguleringsplaner. Staten gjør en rekke ulike typer av investeringer som varierer med hensyn til størrelse, kompleksitet, formål, tidsperspektiv og kompetansebehov. Noen eksempler er:

- Infrastruktur innenfor samferdsel
- Forsvarsmateriell og – installasjoner
- Bygninger
- Energi, infrastruktur og ressursforvaltning
- IT-investeringer
- Statlig hel- eller deleide bedrifter

Vi ser at nye og større datamengder kan være interessant til bruk i både tidligfase av store statlige investeringer som beslutningsgrunnlag (ex ante), og i evaluering av ferdigstilte tiltak (ex post).

Stadig større mengder data blir altså generert gjennom bruk av nye teknologiske systemer og løsninger, såkalte "store data", ofte betegnet som "Big Data". Big Data har både i Norge og internasjonalt blitt det nye begrepet på denne typen data. Vi vil i vår rapport også benytte dette begrepet, men vil innledningsvis drøfte denne betegnelsen og hvorvidt dette egentlig er et nytt fenomen eller ikke. Denne typen data kjennetegnes ved at de inneholder så store og komplekse datasett at de er vanskelig å prosessere gjennom tradisjonelle datahåndteringsverktøy og databaseteknologi. Dette er også data som i mange sammenhenger genereres uten at det foreligger en intensjon om å analysere dataene. Eksempler er Facebooks bildedatabase, pengetransaksjoner, klikk på nettsider, trafikkdata for mobiltelefoner etc. Enorme mengder data kan samles fra mange ulike kilder og aggregeres og analyseres i nye sammenhenger. Det er derimot en rekke utfordringer knyttet til anvendelse av slike store datamengder som for eksempel innhenting, oppdatering, lagring, søk, analysekompetanse, visualisering og personvern.

Big Data omfatter alle de ovenstående punktene. Ut fra kostnadshensyn er sannsynligvis åpne data mest aktuelle å bruke til evaluering. I evalueringsøyemed så er det ikke nødvendigvis viktig med ekstremt store datamengder. Også kvantitative data med omfang som er mindre enn gigabytes og terabytes kan være viktige bidrag til evalueringer, spesielt når de sammenstilles og kombineres med andre typer data. Erfaringene fra innhenting og av analyse av virkelig store datamengder kan derimot være relevant for kvantitative analyser av mindre datamengder også.

Videre i rapporten vil vi i kapittel 2 gi en kort beskrivelse av metode og datagrunnlaget som ligger til grunn for arbeidet med studien. Kapittel 3 ser nærmere på selve fenomenet og begrepet Big Data. I tillegg til å definere begrepet gir vi en oversikt over hvordan Big Data har vært brukt tidligere, innenfor hvilke områder det brukes i dag, samt noen tanker om Big Data i fremtiden. I kapittel 4 setter vi fokus på bruk av Big Data i evaluering av store statlige investeringstiltak. Vi ser nærmere på behovet for data og gir noen eksempler fra andre land hvor man har prøvd å utnytte dette. Kapitlet gir så en drøfting av noen sentrale områder for bruk av Big Data for det formålet vi har fokus på. I kapittel 5 beskriver og drøfter vi muligheter for bruk av Big Data i evaluering innenfor tre relevante prosjektområder: vei, jernbane og bygg. Hensikten er å belyse potensialet for bruk av store data i evaluering av denne type tiltak, samt drøfte utfordringer knyttet til anvendelse av denne typen data. Avslutningsvis oppsummerer vi våre funn gjennom arbeidet med rapporten og gir anbefalinger om videre forskning på området.

2 Metodikk og datagrunnlag

Resultatene som presenteres i denne rapporten er fremkommet gjennom litteraturstudier og samtaler med det vi har ansett som relevante personer i Norge når det gjelder Big Data generelt, og mer spesifikt til bruk i prosjektevaluering.

I litteraturstudien har vi primært gjennomgått litteratur som har omhandlet Big Data, det være seg alt fra populærvitenskapelige artikler, nyhetsoppslag og bransjerelaterte publikasjoner til mer vitenskapelige artikler og publikasjoner. Litteraturen har dekket et bredt spekter av tema deriblant teknologiske løsninger, juridiske forhold, anvendelsesområder, innovasjon mv. Hensikten med gjennomgangen har vært å øke vår egen kunnskap om emnet Big Data, samt kartlegge ulike sider ved temaet som vil være relevant for bruk av Big Data i evaluering av store statlige investeringstiltak.

I tillegg til litteraturstudien har vi gjennomført samtaler med representanter fra ulike miljøer vi mener har relevans for arbeidet med både Big Data som tema, anvendelse i evaluering og caseområdene vi ønsker å beskrive nærmere. Samtalene fulgte noen forhåndsdefinerte tema satt opp i en semistrukturert intervjuguide. Vedlegg A og B viser intervjuguiden samt hvilke miljø som er intervjuet.

3 Big Data

3.1 Hva er Big Data?

Big Data er et forholdsvis nytt begrep som fikk gjennomslag i 2009 (Manyika med flere 2011). Det opprinnelige begrepet som har blitt brukt siden IT-bransjens begynnelse er "data mining" Data mining innebærer at trekker ut informasjon fra store mengder med ustrukturerte data. I prinsippet er Big Data en form for data mining, bare i et svært stort omfang. Det synes ikke å finnes noe etablert norsk begrep for Big Data. Eksempelvis bruker Datatilsynet det engelske utrykket på sine hjemmesider (Datatilsynet 2013). SSB er med i europeisk statistikksamarbeid hvor Big Data er på agendaen. Første utfordring er også her å bli enige om en definisjon av begrepet, i tillegg til å etablere et presist vokabular på området.

En definisjon av Big Data som benyttes av flere, om enn i noen ulike varianter, er at Big Data refererer til *datasett som er så store at de ikke er egnet til å innhente, lagre, prosessere eller analysere ved hjelp av tradisjonelle databaseverktøy* (Nature, 2008; Manyika med flere 2011). Datatilsynet (2013) omtaler Big Data som gigantiske mengder digitale data som er kontrollert av selskap, myndigheter og andre store organisasjoner, og som kan gjøres til gjenstand for omfattende analyse ved bruk av algoritmer

Big Data har noen karakteristika ved seg som gjør at man her snakker om noe annet enn velstrukturerte datasett i en database. Begrepet "the three Vs" som referer til Volume-Velocity-Variety (volum, hastighet, variasjon) er mye brukt (Russom 2011). Et annet kjennetegn er bruk av sanntidsdata.

Det viktigste kjennetegnet som ligger i selve termen, samt oppgis av alle våre intervjuobjekter, er at vi her snakker om data med et stort **volum**. At mengden digitale data med tiden har økt er i og for seg ingen ny trend, men utviklingen de senere årene har vel vært tilnærmet eksplosiv. Et digitalisert samfunn har utviklet seg i rekordfart. Sosiale mediers økende popularitet, samt utviklingen i blant annet sensorteknologi og "Internet of things", har ført til en sterk økning av elektroniske data som også i mange tilfeller ikke er generert av mennesker. Manyika med flere (2011) slår fast i sin rapport at den globale datamengden vokser med 40 % årlig. Datamengdene er så store at de måles i exabytes (en trillion bytes).

I tillegg til store datavolumer er derfor sentrale stikkord hastighet, variasjon og aktualitet. Ved innhenting og bruk av Big Data er det behov for kontinuerlig prosessering ved høy **hastighet**. Avanserte algoritmer som kobler mange ulike, og ofte svært store, datasett skal gjøres simultant. Ofte er det behov for å uthente informasjonen raskt for å ta en beslutning. Dette gjelder eksempelvis for systemer for "drive support" i biler, der data fra ulike sensorer og kameraer hjelper bilførere til å unngå farlige situasjoner. Big Data ble for alvor et begrep da Google utviklet en ny form for modelleringsspråk for behandling av data, såkalt og Hadoop-teknologi. Teknologien gjør det mulig å splitte opp ulike problemer for å kjøre parallelle analyser på store maskinklynger. Prosesseringen går svært rask og man kan hente ut kunnskap i stor skala. Lagringsstrukturene er fleksible og et datasett kan tjene mange ulike behov.

Variasjon i dataene kan på mange måter sees på som selve "gullgruven" med Big Data. Data fra flere ulike kilder blir aggregert og analysert i nye sammenhenger. Potensialet ligger i kobling av data og det å kunne se mønster og trender. Men variasjonen gir også utfordringer når det gjelder prosessering, sammenstilling og analyse. Kostnadene ved å tilrettelegge og analysere data som overhode ikke er innhentet til dette formålet kan fort bli store, og kreve spesiell kompetanse og nye analyseverktøy.

En stor del av de data som karakteriseres som Big Data er **sanntidsdata**. Det er behov for sanntidsinformasjon for å kunne ta raske beslutninger, være seg for å unngå kollisjon, kjøpe aksjer, planlegge etter værmeldingen, eller rekke bussen.

Både omfanget og strukturen på Big Data medfører at man ikke lett kan bruke tradisjonelle former for lagring og analyse av de store datamengdene som for eksempel datavarehus. I stedet kan de som utfører Big Data-analyser kjøpe lagring og prosesseringstid på store maskiner som man får lett tilgang til gjennom

sky-løsninger. Utviklingen av skyteknologi innebærer at eierskapet for datalagring og -analyse er endret. Man får tilgang til maskinvare som prosesserer svært hurtig, og har mulighet til å utnytte og koble sammen flere kilder av data. Det er ikke lengre nødvendig å investere i maskinklynger og hardware. Mange mener at skyteknologi er det som gjør Big Data-analyse i det hele tatt mulig. Flere kommersielle aktører tilbyr i dag slike tjenester som for eksempel Amazon, Google og IBM.

Det er ikke bare tilgangen til dynamiske data som har økt, men også statiske data som digitale kart og bygningsinformasjonsmodeller har blitt vanlige og lett tilgjengelige. Dette innebærer at det nå er lettere å presentere data i en relevant sammenheng. Antall solgte hus og prisen på eiendommene kan eksempelvis vises på et digitalt kart, med mulighet for å søke, zoome og filtrere. I takt med at bruken av digitale bygningsinformasjonsmodeller øker kan ulike informasjon (feilmeldinger etc.) relateres til plasseringen i en bygning. BIM står både for BygningsInformasjonsModell - når man viser til selve modellen og BygningsInformasjonsModellering - når man viser til prosessen for å lage og bruke modellen. Gjennom BIM beskriver man en bygning med bygningsdeler, installasjoner og utstyr som objekter, med definerte egenskaper og relasjoner mellom objektene (Statsbygg 2013). BIM og kart trenger ikke å være Big Data, men de kan brukes til å presentere og analysere Big Data.

Hvor kommer så de store datamengdene fra, og hvilken type data er det egentlig vi snakker om? Hildberg (2013) foreslår følgende inndeling av ulike typer av data og datakilder:

- Tekster (Tracking words)
- Bevegelse og plassering (Tracking locations)
- Omgivelser og miljø (Tracking nature)
- (Internett-) adferd (Tracking behavior)
- Økonomisk aktivitet (Tracking economic activity)
- Annet (Tracking other data)

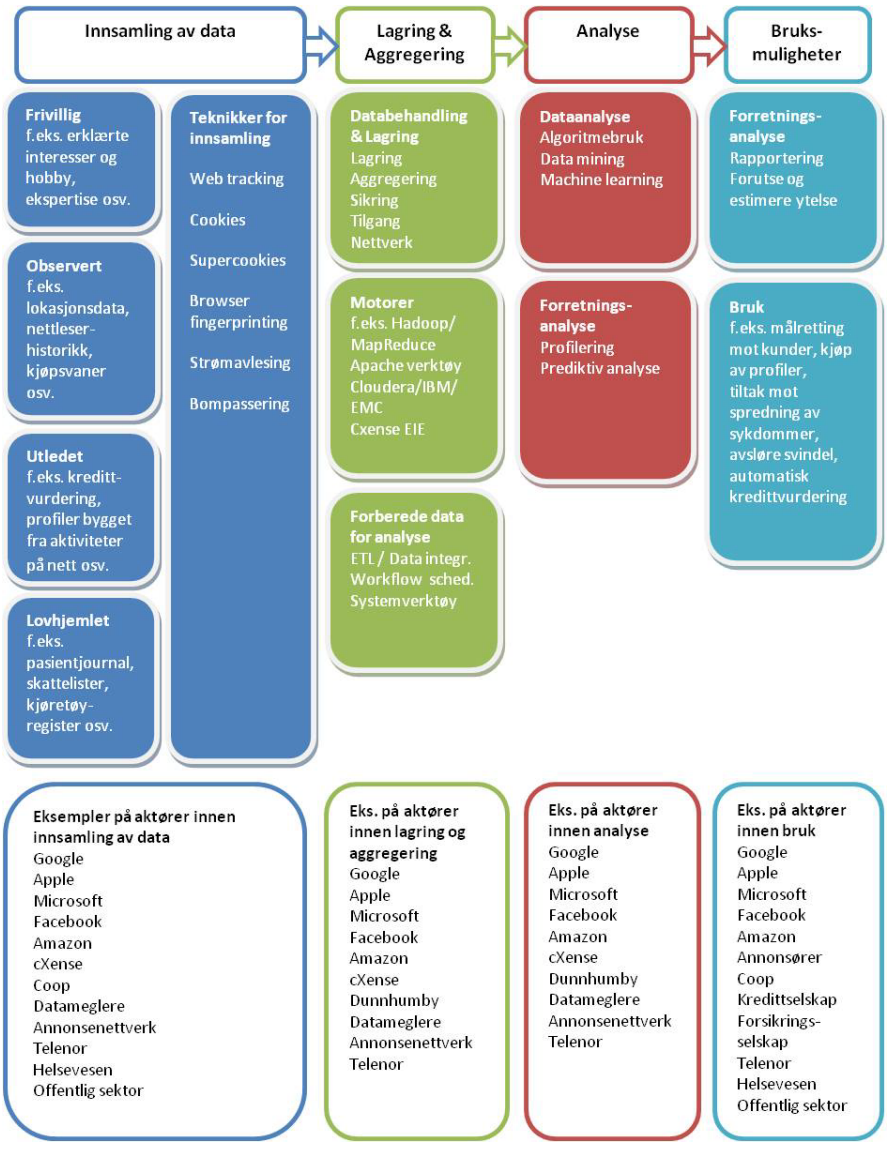
Relatert til evaluering av store statlige investeringer kan oversikten forenkles noe, og vi foreslår en oppdeling i følgende kategorier etter hvordan dataene er samlet inn eller generert:

- Internettrafikk, inkludert aktivitet på sosiale media og data fra søkemotorer (cookies, tekst treff)
- Bevegelsesrelaterte data, inkludert GPS, mobiltrafikk og lokalisering, bomstasjoner
- Data om fysiske omgivelser, inkludert "Internet of things", "machine-to-machine" (M2M) trafikk, typisk fra ulike former av sensorer
- Kommersiell aktivitet, bruk av betalingstjenester, forbruksmønster
- Internt registrerte data i etatene

I Norge har vi store mengder registerdata av god kvalitet som for eksempel befolkningsregisteret, boligregisteret, sysselsettingsregisteret, Norsk Pasientregister mv. SSB mener at dette ikke er Big Data, men andre land mener dette faller inn under definisjonen. Uansett vil registerdata være interessante i tilknytning til evaluering av store statlige investeringer. Det er dessuten viktig å gjøre et skille mellom ferdig

utarbeidet statistikk som nok ikke faller inn under definisjonen av Big Data, og de rådata som SSB sitter på som ligger nærmere Big Data. Det siste kulepunktet trenger ikke å være Big Data siden datamengdene, i hvert fall foreløpig, ikke er så store. I tillegg kan disse data være mer strukturert enn tradisjonelle Big data. Vi tar de med likevel, fordi de er aktuelle å analysere på en måte som ligner Big data, og de er aktuelle å kombinere med Big data. I tillegg har vi sett flere eksempler på at det er muligheter for å bruke disse data mer aktivt til evaluering. Dette gjelder eksempelvis data fra ulike styringssystemer for drift og vedlikehold av offentlig infrastruktur og bygninger.

Datatilsynet gir i sin rapport en god oversikt over ulike innsamlingsmetoder, teknologi, bruksmuligheter og aktører i det de kaller "verdikjeden for Big Data" (Datatilsynet 2013) som vist i Figur 2. Vi mener figuren gir en svært god oppsummering av området.



Figur 2. Verdikjeden for Big Data (Datatilsynet 2013)

Det har vært en rask utvikling innenfor området Big Data de senere år. Følgende viktige utviklingstrekk er verdt å merke seg:

- Større mengder data, inkludert data fra internett og utviklingen av sensor- og sporingsteknologi
- Økt tilgjengelighet til data og press for tilgjengeliggjøring av data
- Tilgang til lagrings- og analysekapasitet gjennom nye løsninger for systemarkitektur og skytjenester med stor og skalerbar kapasitet til en lav kostnad.
- Tilgang til IT-plattformer for å sette data inn i en sammenheng, eksempelvis GIS (geografiske informasjonssystemer) som digitale kart for presentasjon av posisjonsdata, eller bygningsinformasjonsmodeller (BIM)
- Data har fått en egenverdi. Selskaper som besitter store datamengder, spesielt om forbrukeratferd, har i mange tilfeller høyere børsnotering enn selskaper med store realverdier. Selskaper som Wallmart, Facebook og Twitter er eksempler på det.

Det er viktig å skille mellom "Big Data" og "åpne data". Offentliggjøring av data er et viktig virkemiddel for å utnytte potensialet i Big Data. I flere land, inkludert Norge (Difi 2013) og USA (Kali 2012), pågår initiativ for offentliggjøring av datasett for å legge til rette for fremtidig bruk (FAD 2012). I Norge er det lagt til rette for å legge ut datasett på data.norge.no (Difi 2013). FAD opplyser at det er et mål at omfanget av offentliggjøring av data skal øke. Norge er aktivt innenfor OECD for å øke tilgangen på offentlige data. Internasjonalt har Amazon et opplegg for tilgjengeliggjøring av datasett (Amazon 2013). Regjeringen har uttalt at mer data bør være offentlig tilgjengelig (FAD 2102).

3.2 Hva brukes Big Data til?

Spørsmålet er hvordan en kan hente ut innsikt fra et stort datagrunnlag for å kunne ta kunnskapsbaserte avgjørelser (Johansen 2013). Ofte er dataene samlet inn fra flere ulike kilder, for å siden bli aggregert og analysert i nye sammenhenger. Potensialet ligger i kobling av data og det å kunne se mønster, trender og nye sammenhenger. Man kan prosessere nye typer data og utnytte ustrukturert informasjon.

Mest velkjent er kanskje bruken av Big Data i tilknytning til sosiale medier og annen internettbruk. Forslagene på «kjenner du denne personen» som kommer i blant annet Facebook og Linkedin er ofte brukte eksempler på tjenester hvor Big Data-analyse blir brukt. Tilsvarende kartlegges både "klikk" og kjøpemønster på nettet som benyttes til målrettet reklame. Informasjonen er ofte koblet sammen med informasjon om hva andre med ditt kjøpe- eller søkemønster har vært interesserte i, avslørt gjennom deres nettatferd ("klikk"). Ved bruk av Big Data er det mulig å finne etter mønster og sammenhenger det ikke var mulig å få øye på tidligere og å lage profiler på grupper og personer. Big Data brukes i Norge av store aktører i Norge som dagligvarebransjen, mobiltelefonselskaper, energiselskaper og forsikrings- og bankbransjen. Finn.no nevnes også som en aktør innen Big Data da de sitter på store mengder informasjon om kjøpemønstre koblet til blant annet (markeds)priser, boligadresser, telefonnumre og bilder. Også helsevesenet sitter på store mengder data gjennom for eksempel Norsk Pasientregister. Selskaper i oljebransjen har kommet langt når det gjelder generering av Big Data gjennom sensorer på alle tenkelige

driftskomponenter i olje- og gassproduksjon, til å utnytte de til effektivisering av drifts- og vedlikeholdsarbeid.

Sentimentanalyse (tekstanalyse) er et stort område. Man kan søke etter ulike typer uttrykk i tekster, eksempelvis negativt/positivt ladede ord i nyhetsartikler eller Twitter-meldinger, og koble disse mot andre indikatorer (navn på bygg, steder, hendelser etc.). Denne type analyser kan være aktuell for noen type evaluering av store statlige investeringer, eksempelvis hvordan høyprofilerte bygg som Operaen omtales på internett. Mange av bruksområdene for Big Data fokuserer på å gi informasjon i sanntid. Kartlegging av brukermønster til brukere på internett for å tilpasse annonser og andre forslag er et eksempel. Et annet eksempel er beslutningstøtte til offentlige etater. Economist (2013) omtaler hvordan Big Data kan brukes av politiet til forebygging av kriminalitet. Basert på et stort antall ulike datakilder som representerer faktorer som påvirker tid og sted for kriminalitet, kan man peke på plasser med høy sannsynlighet for kriminelle hendelser og styre politiets ressurser dit. Et annet eksempel på bruk av sanntidsdata er omtalt i Harvard Business Review (2012), der data om posisjon og hastighet for fly blir kombinert med data om situasjonen på flyplasser for å gjøre kontinuerlig oppdaterte prognoser for forventet landingstid for fly.

Big Data har også blitt et tema i journalistikken. "Datadrevet journalistikk" er et nytt begrep. Next media er et Arenaprogram[2] som også ser på dette. Flere av dem vi har intervjuet viser til at mediebransjen er interessert i å kunne utnytte Big Data til å frem faktagrunnlag og interessante mønstre, for å belyse eller finne nyheter og interessante tema. Det er stort fokus på dette internasjonalt.

Det private næringslivet har hittil vært lengst fremme i utnyttelsen av den enorme kilden til informasjon som Big Data er. Big Data kan drive frem innovasjon, produktivitet, vekst, nye former for konkurranse og verdiskaping. Potensialet for økt profitt og konkurransefortrinn har trolig drevet frem den teknologiske utviklingen i rekordfart. Her er det store penger å tjene for de som sitter med den nyeste kunnskapen om markeder, kunder og fremtidige trender, og for de som kan utnytte data til økt innovasjon og verdiskaping. Men Big Datas potensiale går utover det å spille en økonomisk rolle for kun det private næringsliv. Manyika med flere (2011) og Hilberg (2013) peker på offentlig sektor som et av de segmenter som har størst potensiale for effektivisering basert på bruk av Big Data.

Manyika med flere (2011) lister flere områder der Big Data kan bidra til verdiskapning og effektivitet:

- Transparens, ved at ulike interessenter (kunder, brukere, velgere etc.) får tilgang til data
- Eksperimentere for å oppdage behov, variasjon og forbedre ytelse
- Segmentering av populasjoner for å skreddersy tiltak
- Erstatte eller understøtte beslutningstaking med automatiserte algoritmer
- Innovasjon i forretningsmodeller, produkter og tjenester

Bruk i evaluering av tiltak kan ses på som et underpunkt under økt transparens.

[2] Arenaprogram er et offentlig virkemiddel for å stimulere innovasjon og samarbeid mellom bedrifter, forsknings- og utdanningsmiljøer og offentlige aktører.

Flere av dem vi har intervjuet fremhever at det er et potensiale til å utnytte Big Data til effektivisering av offentlig sektor. Manyika med flere (2011) lister et antall områder som bør adresseres for å kunne utnytte det fulle potensialet i Big Data. De områder de lister er:

- Rammebetingelser for å kunne utnytte data på tvers av organisatoriske grenser, inkludert avklaring av eiendomsrett, sikkerhet og personvern
- Hensiktsmessige teknologier for lagring og analyse
- Organisatoriske endringer og tilgang på kompetanse
- Tilgang på data

Som en følge av at både privat næringsliv og offentlig sektor ser potensiale for økt effektivisering og verdiskaping gjennom bruk av Big Data, har det vokst fram en erkjennelse om at data er en verdifull eiendel. I følge administrerende direktør Geir Hansen i Geodata, anslås den årlige verdien av offentlige data i EU til å være 140 milliarder euro (Teknisk Ukeblad, 2013). Som en illustrasjon av verdiskapingspotensialet i åpne data nevner han også at etter at GPS ble åpnet for sivilt bruk er det skapt et marked som omsetter for 440 milliarder kroner årlig.

Det bør være et potensiale for bruk av Big Data til offentlig statistikk, og dermed komplettere eller på sikt muligens erstatte tradisjonelle måter å innhente statistikk på. I dag er store deler av datainnsamlingen svært kostnadskrevende og til tider mangelfull når det gjelder innrapportering og svarprosenter. SSB har en pilot hvor de tester ut bruk av transaksjonsdata fra kortbruk til å måle omfanget av grensehandel. SSB opplyser at Big Data på mange måter fortsatt er umodent, også internasjonalt, når det gjelder bruk til statistiske formål. Det vil være nødvendig å tenke nytt når det gjelder statistiske metoder og hva de nye dataene faktisk gir av informasjon. Vil de speile et representativt utvalg av befolkningen? Vil de være av god nok kvalitet til å kunne benyttes til politikkutforming og som beslutningsgrunnlag? Dette er områder som utforskes videre.

Gjennom intervjuene har vi kartlagt hva våre informanter tror og mener om i hvilken retning området for Big Data vil utvikle seg i fremtiden. Noen hovedtrekk er:

- Trenden med offentliggjøring av data vil fortsette.
- Det er en forventning om at data kommer til å få en enda større kommersiell verdi, i tillegg til at volumet av åpne data kommer til å øke.
- Økt tilgang på data samt god tilgjengelighet på datakraft ("store muskler") vil være en driver for innovasjon fremover. De som har dette vil kunne utvikle nye innovative markedsløsninger som gjør dem mer konkurransedyktige.
- Det å eie fasilitetene som dataserverne står i kan bli attraktivt. Norge har her fordeler gjennom tilgang på areal i fjell, kaldt vann til avkjøling, god tilgang på elektrisitet og politisk stabilitet.
- Villigheten til å gå over til skytjenester og eksterne aktører er økende.
- Tradisjonelt flyttes data til de som analyserer, når det gjelder Big Data flyttes analyseprosessene til der data er.

- Det er et stort behov for å utvikle nye metoder for å sammenlikne ustrukturerte data på (eks. utdanning koblet mot helsekøer). SFI i NTNU-regi skal blant annet omhandle dette.
- Big Data mangler metadata, det vil si beskrivelser av datastrukturer (data om data). Dette må det bli mer fokus på fremover.
- En må tenke nytt når det gjelder bruk av statistiske metoder og Big Data. Tradisjonelle tilnærminger til representativitet, signifikans, utvalgskriterier, frafall etc. er ikke så relevant. Nå gjelder det å utvikle metoder for å forstå data som ikke er perfekte og som ikke er samlet inn til et statistisk formål, men som likevel har et potensiale til å brukes.
- Det finnes de som mener at Big Data kan bidra til et trendbrudd for tidligere kvalitativt baserte samfunnsvitenskaper (Morgenbladet 2013; forskning.no 2013). Når man tidligere ofte har vært avhengig av observasjoner, spørreskjema og intervjuer, kan man nå kartlegge folks bevegelser eller internettadferd for store utvalg av mennesker. Dette åpner nye muligheter for studier, som har fellestrekk med naturvitenskapene da de ofte er kvantitative.
- Lage overordnede retningslinjer for eiendomsrett og personvern, også internasjonalt
- Offentlig sektor vil få økt fokus på potensialet i Big Data for å øke konkurransefortrinn og effektivitet. Da vil data også kunne bli lettere tilgjengelig for evaluering.
- Kjøp, salg og videresalg av data vil øke. "Big Data som valuta".

3.3 Big Data og evaluering

Vi har funnet få publikasjoner som eksplisitt omtaler bruk av Big Data til evaluering. Likevel finnes det mange eksempler på bruk av Big Data som kunne vært relevante til evaluering av store statlige investeringer. Dette inkluderer bygg, transport, IT og helse-sektorene. Det finnes sannsynligvis også erfaringer fra bruk av Big Data som er relevant for forsvarstiltak. Som beskrevet innledningsvis vil vi også påpeke at i evalueringsøyemed er det ikke nødvendigvis viktig med ekstremt store datamengder. Også kvantitative data med omfang som er mindre enn gigabytes og terabytes kan være viktige bidrag til evalueringer.

Evalueringer av store statlige tiltak kan adressere resultat, effekt og samfunnsmål. Resultatmål omfatter typisk ferdigstillelsestid, kostnad og oppfyllelse av spesifikasjonen. Effektmål relateres til bruk av prosjektet og samfunnsmål angir de langsiktige og overordnede virkningene av prosjektet. Vedrørende evaluering av tiltak, og store statlige investeringer spesielt, har vi fokusert på muligheter til å bruke Big Data som hjelpemiddel til evaluering av effektmål, og der det synes hensiktsmessig også samfunnsmål. Bakgrunn for dette fokuset er at vi vurderer at det er for disse typer av mål som det har vist seg utfordrende å finne gode data til bruk i evaluering. Data om resultatmål er i de fleste tilfeller intern informasjon som finnes i den utførende etat. Vi kjenner likevel til pågående forskning som undersøker hvordan Big Data kan brukes til operativ oppfølging av tiltak med hensyn på resultatmål, men da typisk med hensikt til å gi prosjektlederen tidlig varsling om eventuelle avvik. I tillegg til evaluering av måloppfyllelse mener vi Big Data kan ha et potensiale når det gjelder å måle ulike tilsiktede og utilsiktede eksterne virkninger av en investering. En kan

også tenke seg at nye datakilder kan avsløre eksterne virkninger man før kanskje ikke hadde muligheten til å observere, gjennom for eksempel kobling av ulike datatyper.

Senseable city lab på MIT (MIT 2013) har brukt Big Data i flere sammenhenger som er interessante for store statlige inverteringer, inkludert transportsystemer, bruk av bygninger og bruk av byrom. De bruker «Big Data» for å beskrive hvordan byer brukes.

Lokasjonsdata, fra eksempelvis mobiltelefon, GPS eller RFID er brukt i flere sektorer for å illustrere bevegelser og adferd til mennesker eller flyt av ulike gjenstander. Dette inkluderer transport (Ferris, Watkins, Borning 2010), turisme (Girardin med flere 2008, Yoshimura med flere 2012) og folks oppførsel i byer (MIT 2013; Arikawa, Konomi og Ohnishi 2008).

4 Sentrale utfordringer ved bruk av Big Data i evaluering

Til tross for at store datamengder genereres hver eneste dag, vet vi at det eksisterer ulike barrierer for utnyttelse av disse data. Dette kan for eksempel skyldes teknologiske vanskeligheter, mangel på informasjon, mangel på forståelse av hva man sitter på av data eller politiske, lovmessige og strategiske hensyn. I dette kapitlet vil vi drøfte ulike utfordringer som alle er relevante for i hvor stor grad man kan utnytte Big Data til evaluering av statlige investeringstiltak i Norge.

4.1 Tilgjengelighet

Å skaffe seg tilgang til data kan i mange tiltak være en tidskrevende oppgave. Både politiske, lovmessige, økonomiske og organisasjonsmessige barrierer kan stå i veien for at man raskt kan få tilgang på de data man trenger. Slik sett er dette med tilgjengelighet et overordnet tema. Flere av de andre områdene vi diskuterer i dette kapittelet vil være relatert til om tilgjengeligheten påvirkes eller ikke, men vi har likevel valgt å diskutere dette separat.

Fra myndighetenes side pågår det i dag et arbeid for å i større grad digitalisere og tilgjengeliggjøre offentlige data. Regjeringen Stoltenberg II ønsket at offentlige datasett skal være åpne og tilgjengelige på nett, for å legge til rette for at næringslivet og andre skal kunne utvikle applikasjoner og tjenester basert på offentlig informasjon. Statlige etater ble i 2011 pålagt å gjøre egnede data tilgjengelig på nett, men mange ligger etter med dette arbeidet (FAD 2012). Det er også laget en veileder for offentliggjøring av data (Difi 2012). Den pågående trenden for offentliggjøring av data bør bidra til å "demokratisere" datatilgangen, slik at det ikke kun er de med store ressurser som får mulighet til å utnytte potensialet i Big Data. Flere av de vi har intervjuet fremhever at det ikke er avklart praksis for hvordan offentlighetsprinsippet i offentlig sektor gjelder for data. Usikkerheten gjelder data som ikke er personopplysninger.

De data som ikke er offentlige kan være vanskelig å få tak i selv om de i utgangspunktet kan utleveres. Private selskaper er ofte mer restriktive pga. konkurransehensyn. Selv om selskapene kan være positive til å bidra med data kan det kreve ressurser å sammenstille og utlevere data, som gjør at tilgjengeligheten blir mindre. Av og til handler vanskelighetene om personvernhensyn og at data må anonymiseres. Dette

diskuterer vi som et eget område litt senere i dette kapittelet. Andre ganger handler det vel så mye om å finne riktige person som har oversikt over de data som finnes. Vårt inntrykk er at mange bedrifter, etater og andre ulike institusjoner selv ikke har oversikt over datamengden de faktisk besitter. Ofte mangler de også kompetansen som trengs for å kunne bearbeide og analysere dataene. Kompetanse diskuteres også som et eget tema senere i dette kapittelet.

Det er gratis å utføre begrensede spørringer mot enkelte sosiale medier. Eksempelvis har Facebook et gratis verktøy man kan laste ned å bruke til å analysere aktivitet på Facebook (Facebook 2013). Verktøyet skal ha innebygde funksjoner som tar hensyn til personvern. En av de vi har intervjuet bruker et verktøy kalt Wisdom fra det amerikansk baserte selskapet MicroStrategy. Han sier: "*Her har jeg tilgang på data fra over 20 millioner mennesker – uten å ha gjort en eneste undersøkelse. Jeg kan analysere ulike preferanser på Facebook, og se på alders- og kjønnsforskjeller mellom ulike grupper og på tvers av land i hele verden. Til nå har jeg sammenlignet kjønnsforskjeller i samfunnsengasjement på Facebook mellom mennesker i Norge, Spania, England, USA, Russland, Egypt, India og Kina".* Her finnes det med andre store muligheter for å skaffe seg store mengde data raskt og til en lav kostnad. Slike typer data er likevel ikke velegnet i alle typer analyser, som vi vil drøfte nærmere i neste avsnitt om anvendbarhet.

Et alternativ kan i noen tilfeller være å selv foreta kartleggingen/registreringen av de data man ønsker. En kan tenke seg at man til evalueringsformål setter opp sensorer som for eksempel fanger opp trafikkmengde på en veistrekning, eller antall personer som besøker et offentlig bygg. Dette kan i noen tilfeller være problematisk med tanke på personvern, samtidig som det vil være en enkel og billig måte å skaffe seg data som da er innhentet til det formålet de skal brukes til. I et evalueringsperspektiv vil det være viktig å kunne sammenlikne den nye situasjonen med situasjonen før tiltaket ble gjennomført. Ved innsamling av egne data må man da helst ha en langsiktig plan knyttet til evalueringen slik at dette blir ivaretatt helst før tiltaket er gjennomført.

Tilrettelegging og analyse av store datamengder er som vi beskrev innledningsvis et stort forretningsområde, men foreløpig rettet mot privat sektor for å blant annet få øket konkurransekraft. Kanskje kan en tenke seg at man i en nær fremtid kan kjøpe ferdig analyserte data av selskaper som spesialiserer seg på dette også for offentlig sektor.

Vi nevnte innledningsvis at en stor del av de data som karakteriseres som Big Data er sanntidsdata, dvs. at de gjøres tilgjengelig umiddelbart etter at de er generert (Google 2013). I tilknytning til evaluering av store statlige investeringer er derimot ikke sanntidsdata nødvendigvis så interessant for evaluator. Man vil typisk ha data som dekker lange tidsperioder, og om mulig data som beskriver situasjonen før et prosjekt starter. Dette kan være mange år tilbake i tiden. Slike data er derimot ofte en mangelvare. For å legge til rette for kommende evalueringer er det ønskelig å iverksette tiltak for å sikre at data blir lagret over lange tidsrom. Det krever at man kjenner hvilke data som kan være aktuelle og at man legger til rette for å gjenfinne disse data når det er tid for evaluering etter flere år. Det er sannsynlig at viljen til å bidra til slik langsiktig datalagring øker om de involverte etatene, eller andre aktører, ser nytten av denne type data. Det er derfor en fordel om dataene også kan gjøre nytte i påvente av kommende evalueringer, eksempelvis som grunnlagsdata for konseptvalgutredninger og andre typer av analyser.

4.2 Anvendbarhet

Big data skaper nye muligheter til å analysere et fenomen basert på ulike typer av data. Dette øker mulighetene for å finne indikatorer som er relevante i forhold til det tiltaket man evaluerer. I evalueringssammenheng kan Big data blant annet anvendes til å:

- Triangulering og kvalitetssikring av data og analyser som inngår i evalueringer
- Komplettere tidligere kvantitative datakilder, forbedre eksisterende evalueringsparametere
- Bidra med nye evalueringsparametere som ikke har vært mulig å inkludere i evalueringer tidligere
- Bidra med kvantitative data på forhold som tidligere har vært basert på kun kvalitative vurderinger
- Illustrere virkninger som ikke har vært mulige å synliggjøre tidligere

Tilgang på flere datasett som belyser samme fenomen kan brukes til triangulering og kvalitetssikring av data og analyser som inngår i evalueringer. Trianguleringen kan inkludere bruk av etablerte typer av informasjon, som intervjuer og dokumentanalyse samt med bruk av for eksempel sensordata fra ulike kilder.

Det finnes flere eksempler der sensordata fra ulike systemer og ulike måleprinsipper kan brukes for å belyse samme fenomen. Vegvesenet har et antall tellepunkter med automatiske telleapparater. Tellinger kan også foretas på basis av registreringer i bomstasjoner. Disse typer av tellinger kan brukes for å komplettere hverandre. I tillegg testes telling basert på blåtann-signaler (i praksis fremst smarttelefoner, men også PC-er og annet elektronisk utstyr). Da en bil kan inneholde blåtann-sendere er det behov for å korrigere blåtann-tellinger mot eksempelvis bomstasjonregistreringer. De vi har snakket med oppgir at det ofte er dyrere å kvalitetssikre målinger enn å utføre selve målingen. Siden data kommer på så mange ulike formater kreves det vanligvis at man bruker mye tid på å "vaske" data, og tilrettelegger de slik at de kan sammenstilles med andre kilder eller benyttes i ulike analyseverktøy og dataprogrammer. Trafikkvolum på en veg kan baseres på ulike typer tellinger men også på vegslitasje. Spor/jevnhet av en vei kan måles med laser og ultralyd, og derved kan man måle slitasje på vei. Det utføres regelmessig målinger etter initialt slitasje. Men den samme type målinger burde kunne brukes til oppfølging av veien også senere. Punktlighet i jernbane kan tilsvarende måles basert på data fra signalanlegg, GPS-data på togene, kjørelogg på togene med hastighetsprofil, data fra dørstengingssystemet og RFID-brikker på gods.

Triangulering og kvalitetssikring av data og analyser kan også gjøres basert på helt ulike typer av data, eksempelvis sensordata kombinert med internettdata. Internettdataene viser hvordan folk eller media omtaler en situasjon. Trafikksituasjonen i London kartlegges kontinuerlig ved bruk av blant annet kameramonitorer. Dette kan kryssjekkes mot innhold og omfang av meldinger i sosiale media for å se hvordan trafikksituasjonen blir omtalt (Kitchin 2013).

Ved at Big Data skaper nye muligheter til å analysere et fenomen basert på ulike typer av data øker mulighetene for en evaluator til å finne indikatorer som er relevante i forhold til det prosjektet man evaluerer. En innvending i denne sammenheng er at risikoen for ”cherry picking” også kan øke ved at evaluerer kan trekke frem de data, eller de tolkninger av et datasett som understøtter et syn på prosjektet. Det finnes en historie om en mann som har mistet sine nøkler, og står under en lyktestolpe og leter. En

dame spør hva som har skjedd, og lurer på hvor han mistet nøklene. Mannen svarer at han tror at han mistet nøklene et stykke unna. – Hvorfor leter du her da, spør damen. – Fordi det er lysere her får hun til svar. Ved ukritisk bruk kan Big Data skape slike opplyste områder, som får oppmerksomhet, uten at de belyser helheten i et evaluert prosjekt. Nå er dette en innvending man kan komme med mot all bruk av kvalitativ og kvantitativ informasjon i evaluering. Riktig brukt kan Big Data bidra til å redusere problemet, ved at man kan bruke flere ulike datakilder. Med referanse til historien skulle man derved kunne få flere lyktestolper og belyse et større område.

Big data kan anvendes til å bedre og komplettere datagrunnlaget på forhold som er inkludert i evalueringer allerede. Dette gjelder eksempelvis tellinger av antall reisende på en veg eller jernbane, eller punktlighet. Formålsbygg i bruk har oftest blitt evaluert basert på kvalitative metoder og Big data gir mulighet til å komplettere dette med kvantitative metoder. Tilgang på nye typer av data bør gi et potensial til å finne nye parametere som kan inkluderes i evalueringer, noe som også kan bidra til å dokumentere virkninger av prosjekter som man tidligere ikke har klart å synliggjøre. Eksempel på nye parametere er ulike former for internettbruk som kan vise hvordan et ferdigstilt prosjekt brukes og omtales.

Det vil være ulike behov for data ved evaluering av et investeringsprosjekt i tidligfase sammenliknet med en ex post evaluering av samme prosjekt. Ulike evalueringsmetoder og kriterier vil også kunne etterspørre ulike typer data. For eksempel kan evalueringsobjektet ha både operasjonelle og strategiske mål som skal evalueres. Spesielt interessant vil det være å se om denne type data kan avdekke eksterne virkninger som tidligere ikke har vært like lett å kartlegge eller vært ukjente.

Innenfor samfunnsøkonomifaget er det et ønske om å synliggjøre økonomiske ringvirkninger av større prosjekter. Big data bør kunne bidra til denne synliggjøringen. Det hadde vært spennende å se om data over korttransaksjoner kan bidra til å synliggjøre ringvirkninger for kommersiell aktivitet av for eksempel et transporttiltak.

Big Data har også potensiale til å erstatte eller komplettere, andre typer av undersøkelser. Studier av betalingsvillighet innfor transport har de seneste 10-årene for en stor del vært basert på spørsmål til brukere, så kalte "stated preferences" (SP)-studier. En alternativ type undersøkelse er å studere faktiske valg av transportform og betaling, så kalte "revealed preferences" (RP)-studier. Det har tidligere vært vanskelig å få data til RP-studer, noe som har medført at SP har vært de vanligste. Dette kan endres dersom store mengder data om faktisk utførte reiser blir tilgjengelig. I forlengelsen kan Big Data eventuelt bidra til at det blir lettere å prissette nytte i en nytte-kostnadsanalyse innen andre områder enn transport.

4.3 Relevans

Big Data er gjerne samlet inn på en uvanlig måte for en statistiker. Det meste av statistisk teori og analyse er basert på at man studerer en populasjon ved å se på et tilfeldig utvalg fra en populasjon. Innen Big Data segmentet er det ofte at data ikke er et tilfeldig utvalg, men samlet inn slik at innsamlingsmetoden må være med i modellen for hvordan data skal analyseres. Dette medfører at man trenger nye statistiske metoder for å forstå data som ikke er perfekte og som ikke er samlet inn til et statistisk formål, men som

likevel har et potensiale til å brukes. Man må tenke nytt når det gjelder bruk av vitenskapelige metoder. Tradisjonelle statistiske problemstillinger som representativitet, signifikans, utvalgskriterier, frafall etc. må tilpasses de nye typene av data. Vi tror at disse begrepene fortsatt vil være relevante, men tror at endring i vitenskapelige metoder vil bli mer aktuelt når Big Data-analyser blir mer etablerte. Et spørsmål er om bruk av Big Data i fremtidens beslutningsprosesser i det hele tatt kan sammenliknes med bruk av mer tradisjonelle statistiske kilder. Trolig vil Big Data i mange sammenhenger være et supplement til de tradisjonelle kildene. For å sikre relevans og pålitelighet ved bruk av ulike Big Data-kilder bør kildene kvalitetssikres gjennom for eksempel triangulering mot andre mer tradisjonelle datakilder.

En av fordelene med Big Data som ofte fremheves er mulighetene til å gi informasjon i sanntid. I forhold til bruk av Big Data i evalueringer (og kanskje spesielt ex post) synes det ikke å være viktig at analysene skjer i sanntid. Men funksjonalitet som kontinuerlig gir opplysninger om eksempelvis forventet kjøretid på en vegstrekning gir nyttig informasjon til brukerne av vegen, og kan også ses på som en form for evaluering av tiltak som gjennomføres på vegen.

En annen utfordring er at data bare samles inn om aktiviteter der en gitt type teknologi er i bruk. Dette kan være en vesentlig begrensning i forhold til representativitet av data i noen tilfeller. Eksempelvis varierer bruken av sosiale media, bruk av mobiltelefon og det er ulik funksjonalitet på mobiltelefonene. Endring i bruk henger sammen med at brukermassen også har endret seg. Facebook ble for eksempel i sin tidlige fase i hovedsak benyttet av ungdom, mens vi ser at mediet i dag benyttes av både gamle og unge. Sensordata og data fra kommersielle transaksjoner synes å være mer sammenlignbare over tid. Posisjonsdata kan være påvirket av hvilken teknologi som brukes for å registrere posisjon og bevegelse. Utfasing ev en teknologiplattform (som det gamle mobiltelefonsystemet NMT900) og innfasing av en annen (som smarttelefoner, eller telefoner med blåtann) kan skape utfordringer ved sammenligning av data over lange tidsperioder. Dette innebærer at analyser basert på Big Data kan være mer relevante til å beskrive situasjonen ved tidspunktet for evaluering, sammenlignet med å beskrive utviklingen over en lengre periode. Folks adferd kan ha blitt endret i løpet av prosjektperioden. Dette kan skape en utfordring ved bruk i ex post-evaluering. Data fra tidligfaseanalyser eller om situasjonen før prosjektet er ikke fullt ut sammenlignbare med ex post-data. Mange store statlige investeringer løper over flere år. For sensordata kan måleprinsippet, lagringsmåten eller aggregeringsformatet endres over tid. Disse utfordringene kan reduseres dersom data lagres med høyest mulig oppløsning, og det angis tydelig hvordan dataene er innsamlet og bearbeidet. Det er likevel sannsynlig at data som samles inn i dag kan sammenliknes direkte med data fra samme kilde om 10 år, i den grad de samme kildene er i bruk om 10 år. Dette gjelder spesielt data om fysiske omgivelser som måler uten noe særlig utvalgskriterium, eksempelvis værdata.

For å kunne sikre sammenliknbare data over tid må det iverksettes tiltak for å sørge for at data blir lagret over lengre tidsrom allerede når man starter planleggingen av et tiltak. Det krever at man kjenner hvilke data som kan være aktuelle, og at man legger til rette for å gjenfinne disse data når det er tid for evaluering etter flere år.

4.4 Personvern

Big Data utfordrer sentrale personvernprinsipper. Personvernhensyn vil i mange tilfeller være en barriere når det gjelder tilgjengelighet og anvendelse av disse data, men trenger likevel ikke å være en fundamental hindring slik mange tror. Også Datatilsynet påpeker at til tross for flere personvernutfordringer, er det mulig å benytte Big Data på en måte som ivaretar den enkeltes personvern.

Datatilsynet har fokus på Big Data og følger utviklingen nøye. I september 2013 publiserte de rapporten "Big Data – Personvernprinsipper under press". Her trekker de frem ti sentrale personvernutfordringer knyttet til Big Data og diskuterer personvernutfordringene innenfor ulike anvendelsesområder som internettbaserte selskaper, forsikring, kredittvurdering, helse og politi. De foretar også en gjennomgang av både norsk og internasjonalt lovverk på området.

I Norge har man ikke noe lovverk som eksplisitt omhandler datasikkerhet i forbindelse med bruk og analyse av Big Data. Det finnes lite norsk og europeisk rettskildemateriale som direkte berører tematikken (Datatilsynet 2013). Utgangspunktet for de spørsmål som berører personvern vil være Personopplysningsloven[3]. Loven gjelder alle former for personlig informasjon som kan knyttes til en identifiserbar enkeltperson. Her vil det likevel i praksis kunne være rom for mange tolkninger av både tilknytningskravet og i hvor stor grad personene er identifiserbare. Når det gjelder innsamling og analyse av Big Data med personvernopplysninger står to rettslige grunnlag sentralt i personopplysningsloven: *lovhjemmel* og *krav om samtykke*. Dette sikrer at data enten skal samles inn med hjemmel i lov (for eksempel gjelder statistikkloven for SSB), eller at det skal være innhentet samtykke fra alle de som opplysningene angår. Kravene om hjemmel i lov eller samtykke vil være relevante i de tilfeller hvor en i forbindelse med en evaluering vil ønske å selv innhente informasjon. Kravene begrenser og fordyrer trolig mulighetene når det gjelder innhenting men trenger ikke å være en absolutt hindring for innsamling av data. Datatilsynet diskuterer også i sin rapport muligheter for å benytte systemer hvor personvernhensyn er innebygd i alle deler av databehandlingen, fra innsamling til bearbeiding og analyse.

Kravene i personopplysningsloven, gjelder bare dersom opplysningene kan knyttes til identifiserbare personer, og ikke behandling av anonyme opplysninger. Ved utlevering og behandling av data kan personvern derfor likevel overholdes gjennom aggregering og anonymisering av data. Analyser basert på kun et datasett er i utgangspunktet forholdsvis takknemlige å aggregere slik at de ikke gir personopplysninger. Gjennom sammenstilling av data fra flere kilder kan det likevel oppstå risiko for at enkeltindivider kan identifiseres fra i utgangspunktet anonyme datasett (Narayanan og Shmatikov 2009). Spesielt vil en kobling mot data på svært detaljert geografisk nivå i mange tilfeller gjør det enkelt å identifisere personer. For eksempel kan anonymiserte data om et kjøretøys bevegelser kobles mot lokasjonsdata fra en mobiltelefon, og det vil dermed være mulig å sannsynliggjøre hvem sjåføren er. For bruken av Big Data blir dette et dilemma, da det nettopp er koblingen av mange ulike datakilder som gir store analysemuligheter. Bedre rutiner og teknikker for anonymisering av data er derfor et viktig område

[3] Personopplysningsloven, lov av 14. april 2000 nr. 31 om behandling av personopplysninger, bygger på det europeiske personverndirektivet (Europaparlamentets og Rådets direktiv 95/46/EF av 24. oktober 1995 om beskyttelse av fysiske personer i forbindelse med behandling av personopplysninger og om fri utveksling av slike opplysninger).

fremover. Relevansen av data og analyser kan også bli utfordret ved bruk av aggregerte data for å ivareta personvernhensyn. Dersom variasjonen innenfor hver gruppe i aggregatene blir for stor kan man stille spørsmålstegn ved relevansen av resultatene i analysen.

En annen utfordring for data med personopplysninger er at de ikke kan benyttes til andre formål enn det de var samlet inn til. Dette hindrer muligheten for gjenbruk av data og begrenser selvsagt analysemulighetene. Et relevant eksempel i vårt tilfelle vil være hvis man hadde ønsket å registrere hvor mange som hadde vært på operataket ved bruk av lokasjonsdata fra for eksempel mobiltelefoner. Ved å analysere de ulike brukernes lokasjonsdata kunne man i prinsippet også avdekket hvilke andre turistattraksjoner de hadde besøkt den samme dagen. Dette kunne ha vært av interesse for å få et bredere perspektiv i analysen, men er utfordrende når det gjelder personvern, og medfører sannsynligvis et krav om samtykke for innhenting av data. Hvis formålet med datainnsamlingen var å kartlegge turister på operataket kan dataene ikke brukes til å si noe om andre forhold. Mangel på åpenhet og informasjon om hvordan data benyttes og sammenstilles er et annet område også Datatilsynet adresserer. Siden bruk av disse data i privat sektor ofte dreier seg om å øke konkurransekraft og skaffe seg fortinn i ulike markeder, ser man en mangel på transparens rundt bruk og analyse av Big Data.

Et hovedinntrykk vi sitter igjen med etter intervjuene er at den teknologiske utviklingen løper i fra lovverket på dette området. Temaet er absolutt på dagsorden både i Norge og internasjonalt, og det er satt i gang en rekke initiativ for å bedre situasjonen. EU-kommisjonen arbeider for tiden med en revidert forordning om behandling av personopplysninger. Forordningen vil bli gjeldende rett for Norge, ettersom lovgivningen på databeskyttelses-området er EØS-relevant. Også i USA jobbes det med å bedre lovgivningen for data-sikkerhet og mer spesifikt personvernhensyn og Big Data. Mange av aktørene (sosial medier, analysebedrifter, kommersielle aktører) er amerikanske så man kan tenke seg at lovgivningen som utvikles i USA vil kunne legge føringer også i andre land.

I tråd med Datatilsynets konklusjon trenger altså ikke personvernhensynet nødvendigvis være en hindring for bruk og analyse av Big Data. Lovgivningen gjør likevel at man støter på ulike utfordringer som kan gjøre tilgjengeligheten mindre og anvendelsen noe begrenset.

4.5 Eiendomsrett

Eiendomsrett er nært beslektet med temaet personvern når det gjelder bruk av Big Data, men omhandler likevel litt andre forhold. Med eiendomsrett tenker vi her på hvem som faktisk sitter på rettighetene til informasjonen som samles inn, være seg med eller uten personopplysninger. To gjeldende prinsipper synes å være at den som samlet inn dataene eier dem, og at den aggregerte versjonen av data eies av den som har utført aggregeringen. Et annet viktig tema i dag og spesielt fremover vil være sporbarhet i bruken av data. Hvem har brukt de, hvem har hatt innsyn? Det vil i fremtiden bli viktig å kunne ha gode og pålitelige systemer rundt dette.

Bruk av Big Data i kommersielle sammenhenger de senere år, har gjort at flere har fått øynene opp for hvor verdifullt det er å sitte på store mengder data som kan gi ny og forbedret innsikt og kunnskap. Private

selskaper henter ut store verdier ved å analysere dataene, mens vi som gir de i fra oss ser lite til dette. Dataene i seg selv er blitt verdifulle og dermed blir det også viktig å avklare eiendomsrettighetene.

Vårt inntrykk gjennom arbeidet med denne studien er at det hersker usikkerhet rundt dette med eiendomsrett og Big Data. Dette gjelder særlig data som er automatisk genererte gjennom for eksempel bruk av sosiale medier eller sensorer. Mercur kjøpesenter i Trondheim tilbyr gratis wifi til sine kunder. De har derfor tilgang på informasjon om alle enheter som er pålogget deres lokale wifi-nett. Men eier de også da disse data og står de for eksempel fritt til å selge de videre? I utgangspunktet er svaret på dette ja. De som samler inn, evt. anonymiserer eller aggregerer dataene eier også dataene.

Offentlig sektor har som nevnt begynt å forstå at store data har en verdi, dette kan endre regler både for personvernhensynet og eiendomsrett. Spørsmål om eiendomsrett kan også påvirke tilgjengelighet og kostnader. Tilgjengelighet kan bli redusert når organisasjoner som har interessante data ikke ønsker å dele dem. Dette er for eksempel tilfelle for høyoppløselige data relatert til antall reisende på tog. Av forretningsmessige hensyn ønsker ikke togoperatørene at disse data skal være offentlig tilgjengelige. Men de er åpenbart relevante ved evaluering av investeringer i jernbanespor og annen infrastruktur. Kostnader kan påløpe dersom organisasjoner, kommersielle eller offentlige, ønsker å ta betalt for å gjøre data tilgjengelig. Kommersielle aktører kan gjøre dette av kommersielle grunner. Offentlige organisasjoner kan gjøre det for å få dekket egne kostander til datainnsamling og bearbeiding, eksempelvis aggregering og anonymisering. I begge tilfeller skulle det medføre at bruk av de aktuelle data blir dyrere og derved blir brukt i mindre omfang. Samfunnet kan derved gå glipp av grunnlagsinformasjon om bruk av betydelige mengder av våre skattemidler. Vi håper at diskusjoner om eiendomsrett ikke skal bli en hindring for bruk av offentlige organisasjoner sine data, og at disse kan brukes til evaluering.

4.6 Kostnad

Evalueringsoppdrag er ofte begrenset i tid og ressurser. Kostnader knyttet til innhenting og analyse av data utgjør ofte en stor del av rammene i prosjektene. Spesielt kan datainnsamlingen være både dyr og tidkrevende. Det vil derfor være relevant å vurdere kostandene ved bruk av ulike datakilder, og veie disse opp mot nytten man har av disse dataene. Et viktig aspekt er om bruk av Big Data erstatter tidligere kilder, eller kommer i tillegg. Kommer de i tillegg vil de samlede kostnadene trolig øke, men samtidig kan nytten av å få disse data likevel være stor nok til å kunne forsvare kostnaden. Kanskje kan de for eksempel brukes til å måle effekten av ulike tiltak som man før ikke hadde mulighet til å måle. Dersom Big Data kan erstatte mer tid- og kostnadskrevende innhenting av data til evaluering så oppnår man en effektivitetsøkning. Denne kan tas ut i mer evalueringer, eksempelvis evaluere flere tiltak, eller redusere kostnaden for evaluering, i beste fall begge deler.

Det er særlig innhenting og prosesseringen av data som ser ut til å være mindre ressurskrevende når det gjelder Big Data generelt. Utvikling av Hadoop-arkitektur og skyløsninger gjør enorme mengder data tilgjengelig med et tastetrykk, og reduserer prosesseringstiden betraktelig. Lagring av data er også i mange tilfeller lite kostbart. Tilsynelatende er det minimale "upfront" investeringer, siden man ikke trenger å

investere i maskincluster etc. Ved bruk av sky-teknologi har man dessuten hele tiden tilgang på de mest oppdaterte teknologiske løsningene, uten selv å ha investert i dette.

Til bruk for statistiske formål vil Big Data kunne lette oppgaveplikten til selskaper som må rapportere data, samt redusere behov for skreddersydde surveys som krever mange ressurser. For eksempel kan bruk av kortdata avsløre forbruksmønster som kan erstatte (deler av) den svært ressurskrevende forbruksundersøkelsen. Myndighetenes satsing på mer åpne data vil kunne gjøre datafangsten billigere.

Nye teknologiske løsninger har ført til at det heller ikke nødvendigvis trenger å være så dyrt å samle inn egne data. Teknologien er billig (sensorer er for eksempel hyllevare) og man kan samle inn på det formatet man vil ha. Dette kan redusere ressursbruk på vasking og tilrettelegging av data. Det diskuteres også hvor billig løsningene faktisk er for den enkelte bedrift som ønsker å benytte seg av Big Data-teknologiene. Mange Web-selskaper som kjører Hadoop stoler helt på redundans av data, men hvis du er en bedrift, bank eller en statlig etat, må standarder for databehandling følges, med blant annet rutiner for sikkerhet, gjenoppretting og tilgjengelighet. Dette introduserer en mer komplisert forvaltning og ikke minst behov for kompetanse på området. Kompetanse for analyse og prosessering av data er også nødvendig, og er en kostnadskomponent i seg selv.

4.7 Kompetanse

Flere store internasjonale aktører og fagmiljøer innen Big Data har pekt på at det å skaffe kvalifiserte folk med riktig kompetanse innen databehandling og analyse er en flaskehals for utviklingen innen området (Gartner, Mc Kinsey, The Guardian). Gartner Inc. konkluderte allerede i 2012 at utviklingen innen Big Data ville skape behov for 4,4 millioner nye IT-jobber på verdensbasis frem mot 2015. Selskapet anslo at bare en tredjedel av jobbene ville bli fylt med kvalifisert personell.

Big Data krever ulike typer kompetanse. I tillegg til IT-kompetanse som artifisiell intelligens, trengs også svært god kunnskap på analyse og visualisering av data. Bruk av artifisiell intelligens innen utvikling av maskinlæringsteknikker[4] kan muligens til dels kompensere for eventuell mangel på analysekompetanse, men dette ligger litt frem i tid. Behovet for kompetanse krever også en satsning innen undervisningssektoren som bør legge til rette utdanningstilbud for å møte behovet. Området krever også forskningskompetanse, både for å forstå hvordan man rent teknisk kan prosessere data i tillegg til hvilken kunnskap man kan hente ut av dem.

Flere av intervjuobjektene mener at Norge ikke er langt fremme når det gjelder kompetanse på analyse og prosessering av data, selv om flere IT-miljøer (Telenor) har begynt å bygge kompetanse på området den siste tiden. Det har inntil nå vært mest fokus på tradisjonell statistikk, datavarehus og rapportering. Mange leier inn konsulenter, for eksempel varehandelsbedriftene, for å gjøre store analysejobber. Svært få aktører kjenner teknologien (hadoop) så de kan utnytte den selv. Dette gjør dem avhengige av ekstern ekspertise. Manglende kompetanse kan derfor også være en barriere for utnyttelsen av Big Data i Norge. For å løse dette er det viktig at både offentlig sektor og privat næringsliv er klar over denne utfordringen, og er i

[4] Analyseresultatene benyttes som informasjon for å gjøre nye analyser bedre.

forkant av behovene når de oppstår både ved å tilrettelegge for utdanning og gjennom målrettet rekruttering.

5 Utvalgte case-områder

I det følgende diskuterer vi muligheter for bruk av Big Data til oppfølging av transport- og byggetiltak. Transportsektoren står for en stor andel av de prosjektene som kommer under KS-ordningen. Det er også en sektor der det er prøvd ut det vi kan kalle Big Data-inspirerte datainnhentingsmetoder, og som synes å ha et potensiale for utvidet bruk av nye typer av data. Offentlige bygninger er en type store statlige investeringer som av tradisjon er evaluert basert på kvalitative metoder. Det synes å være et potensiale for å komplettere denne type evaluering med bruk av Big Data. En forholdsvis stor prosjektgruppe som vi ikke tar opp her er forsvarsprosjekter. Det er grunn til å anta at det finnes mye interessante data relatert til forsvarsaktivitet, men mye av disse data er høyst sannsynlig gradert. Nettopp derfor kan kanskje bruk av Big Data ha et potensiale. Det er få IT-prosjekter som er så store at de er underlagt KS-regimet, og vi har valgt å heller ikke se nærmere på denne typen prosjekter når det gjelder bruk av Big Data.

5.1 Samferdsel

Big Data åpner mange muligheter for analyser av samferdselstiltak. De mest åpenbare mulighetene ligger i å måle trafikkstrømmer på nye og utvidede måter. Det mest spennende er mulighetene til å kunne se på transportmønster og ikke bare måle trafikkvolum ved de punkter der det finnes en telling. Man kan bruke ulike former for trafikkmålinger i kombinasjon for å kvalitetssikre de ulike datakildene. Man kan også søke forklaringsfaktorer ved å kombinere trafikkdata med eksempelvis værdata.

En interessant form for måling er reisetid. Det kan være reise fra hjemmet til arbeid. Å måle endring i reisetid fra hjem til arbeid (og omvendt) og endring i reisemønster før og etter en større investering (motorvei, nytt jernbanespor etc.) for en større befolkningsgruppe, hadde vært en god måte å evaluere effekten av investeringstiltaket på. Big Data åpner muligheten for den type evaluering. Og det synes mulig å gjøre det uten å komme i konflikt med personvernhensyn.

Calabrese med flere (2012) viser at mobiltelefondata kan brukes til å beskrive folks bevegelsesmønster, som et alternativ til de tradisjonelle transportkartleggingene. De brukte mobiltelefonregistreringer fra en million brukere i Boston i løpet av tre måneder, for å beskrive transportbehovet. Tilsvarende studie bør være mulig å gjøre før og etter ferdigstillelse av større investeringer i infrastruktur. Calabrese peker på tre utfordringer ved bruk av mobildata: (1) økonomiske og demografiske opplysninger om enkeltpersoner er ikke tilgjengelig grunnet personvernhensyn (2) mobilbrukere er ikke nødvendigvis representative for hele befolkningen og (3) data er ikke formaterte for denne type analyser. For å adressere den første utfordringen brukte de aggregerte data der brukerne ble samlet i grupper som tilsvarer det mest detaljerte nivået som økonomiske og demografiske data var tilgengelige. Til å kalibrere og kvalitetssikre dataene ble det brukt informasjon fra sikkerhetsinspeksjoner av kjøretøyene, som inkluderte km-stand.

Telenors forskningssenter oppgir at de undersøker mulighetene for å utføre lignende type analyser basert på norske mobiltelefondata. Vi tror det hadde vært interessant for Concept å etablere et samarbeid med dem for eksempel i forbindelse med å utføre en pilot på bruk av Big Data i evaluering av transporttiltak.

Big Data gir allerede i dag mulighet til å innhente store mengder data av bilpasseringer. Ulike former for tellinger av bilpasseringer oppgis å være forholdsvis lett og billig å innhente. Det finnes 2000 faste tellepunkt i Norge, og det er mulig å plassere ut ytterligere målere. Nøyaktigheten på tellinger er bedre for en bomstasjon enn vanlige tellepunkt. Bomstasjonsmålinger kan derfor brukes for kontroll og kalibrering av de andre målingene. I tillegg pågår forsøk med å bruke tellinger basert på deteksjon av blåtann-enheter som passerer etter vegen.

Andre data enn tellinger og mobiltelefonregisteringer kan være aktuelt å benytte. For å måle trafikkvolum kan man med stor nøyaktighet måle slitasje på vegbanen ved hjelp av laser og ultralyd. Dette gjøres regelmessig på nye veger, men burde også kunne gjøres senere. Volum for både bil- og flytrafikk kan også registres ved videokameraer.

Andre typer av data enn de relatert til trafikkvolum er også aktuelt. Dette gjelder dels mulige forklaringsfaktorer, som værdata, og data om hvordan trafikken oppfattes, som omfanget av Twitter-meldinger for eksempel på temaet «sitter fast i kø på E18».

Tabell 1 er en illustrasjon av indikatorer, kilder og barrierer relatert til bruk av nye data i evaluering av vegprosjekter. Tabellen, og senere tabell 2 og 3, tar utgangspunkt i generelle effektmål for store statlige investeringer innenfor samferdsel. Hvert prosjekt har sine egne unike mål, men det er likevel en del elementer som erfaringsvis går igjen i denne type prosjekter. De generelle målene brukes her til å illustrere hvordan Big Data kan brukes til å lage indikatorer som kan brukes til å måle måloppnåelse.

Innenfor samferdsel (veg og jernbane) er de generelle effektmålene relatert til spesielt tid og volum. Tid er ofte relatert til reisetidsreduksjoner, men kan også være koblet til reduksjon av variasjon i reisetid, det vil si økt forutsigbarhet, fremkommelighet eller punktlighet. Volum innebærer eksempelvis antall kjøretøy (ÅDT), antall reisende på toget eller godsvolum som transporteres. I tillegg kan samferdselsprosjekter ha mål relatert til sikkerhet, miljø og økonomisk utvikling.

Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
Internett-aktivitet	(opplevd) tid	Brukertilfredshet, tidsbesparelser	tekstanalyse av medieoppslag	God	Relevant for å komplettere bildet. Utfordringer knyttet til signifikans, endring over tid, tekniske fremskritt.	Ingen, offentlige kilder	Middels. Software til analyse er hyllevare.
	(opplevd) tid	Brukertilfredshet, tidsbesparelser	Analyse av sosiale media, Twittermeldinger, facebook	God	Relevant for å komplettere bildet. Utfordringer knyttet til signifikans, endring over tid, tekniske fremskritt.	Basert på åpent tilgjengelig informasjon	Middels. Software til analyse er hyllevare.
Bevegelser	Volum	ÅDT, antall biler	Bomstasjon, tellepunkt	Krever at data gjøres tilgjengelig	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	
	Tid	Tidsbesparelser	Bomstasjondata, tid mellom passeringer av to eller flere bommer	Krever at data gjøres tilgjengelig	Måler en viktig effekt, men med få tellepunkter	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggs-kostnad fremst på sammenstilling og analyse
	Tid	Tidsbesparelser	Mobitefondata, fra basestasjon hjemme til basestasjon arbeid	Krever at data gjøres tilgjengelig. Tilnærmet kontinuerlig oppfølging	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggs-kostnad fremst på sammenstilling og analyse
Fysiske omgivelser	Volum	ÅDT, antall biler	Vegslitasje	Kan måles av Vegvesenet	Kompletterer andre kilder	Ikke personopplysning	Avhenger av Vegvesenet
	Volum	Kjørelengde biler til service og/eller EU-kontroll	Verksteder	Krever at data gjøres tilgjengelig	Viser aggregert kjørelengde. Kan vise kjørelengde for enkeltbiler også	Avhenger av aggregering	Usikkert
	Volum	Luftkvalitet, støy, etc.	Ulike former for målere	Sannsynligvis god, da det typisk er offentlige (kommuner, vegvesen etc som måler)	Måler volum indirekte, kompletterer andre data	Ikke personopplysning	Forhåpentligvis gratis tilgjengelig
Kommersiell aktivitet	Volum, eksterne effekter	Ringvirkninger	Betalingskort	Finansinstitusjoner	Måler på individnivå eller aggregert	Avhenger av aggregering	Må evt. kjøpes i et kommersielt marked

Tabell 1. Illustrasjon av mulige indikatorer, datakilder og bruk av nye typer data til evaluering av vegprosjekter. Merk at vurderingene om eksempelvis kostnad og personopplysning er foreløpige. (Se vedlegg C for en større versjon).

Når det gjelder jernbane, er dette et tett integrert system, der alle aktørene er avhengige av hverandre. Skinnegangen tillater mye mindre fleksibilitet enn hva eksempelvis veier, luftrommet og sjøen gir. Alle bevegelser på jernbanen må derfor til stor grad planlegges, og trafikken må overvåkes og koordineres kontinuerlig. Jernbanen er blitt beskrevet som en stor maskin, til forskjell fra de fleste andre transportformer, der hvert kjøretøy, fly, båt og de ulike delene i infrastrukturen er ”egne maskiner”. Fordi jernbanen har en egen infrastruktur, er jernbanen forholdsvis isolert fra andre transportformer rent driftsmessig. At jernbanen er så isolert, innebærer at det ligger til rette for å få forholdsvis god oversikt over hvilke data som finnes og hvem som sitter på dem. Det isolerte systemet kan også ha medført at data oppfattes som organisasjonsinterne, og at det ikke er så stor tradisjon for å inkludere data relatert til forhold utenfor systemet.

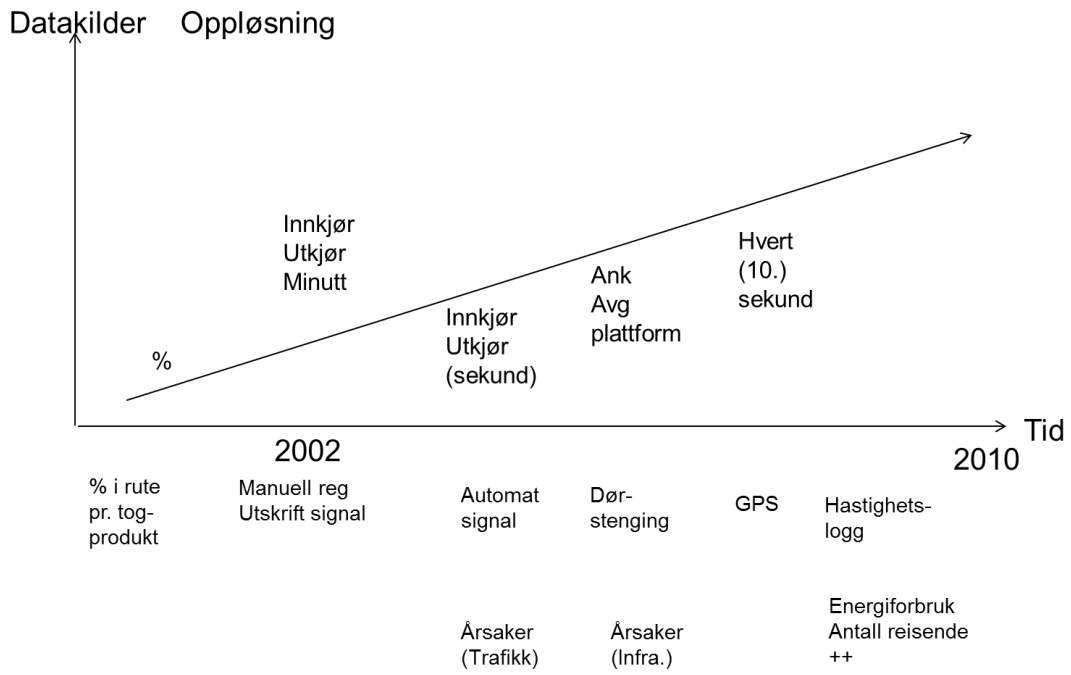
Aktuelle data relatert til nytten av investeringer i nye jernbaneinfrastruktur inkluderer reisetid, antall tog, punktlighet og antall reisende. Dette er data som på aggregert nivå har vært mulige å finne i lang tid (se Olsson 2006). Nye datakilder og analysemuligheter åpner for bruk av flere ulike datasett og mer høyoppløselige data. Dette kan illustreres med utviklingen i tilgang på punktlighetsdata.

Inntil cirka 2003 var punktlighetsdata i praksis kun tilgjengelige som prosent tog i rute til endestasjon. Dataene bar basert på manuelle registreringer og på utskrifter fra et gammelt Norsk Datasystem. Da systemet var gammelt og ikke kunne oppgraderes var det ikke lagringsplass for dataene. For å få historikk måtte man spare papirutskriftene. I tillegg ble det gjort manuelle notater av togledelsen i de områder der det ikke var automatisk togstyring, blant annet nord for Trondheim. For å gjøre analyser av punktlighet og årsaker til forsinkelser måtte man innhente grunndata på papir og registrere de i eksempelvis Excel.

Ved innføring av systemet TIOS fra 2003 ble togbevegelser automatisk lagret og det ble mulig å analysere hvert tog sine bevegelser, basert på når togene passerte kjøresignalene. I tillegg blir årsaker til forsinkelser registrert.

Nye tog, fra Signatur/Flytoget som kom i 1998, har i økende grad ulike systemer som lagrer data fra sensorer i toget. Spesielt når disse data kan kobles til GPS-registreringer gir de mulighet til å kartlegge togenes bevegelse mer nøyaktig enn hva som er tilfelle når man kun bruker data fra signalsystemene. Data som er interessante inkluderer GPS-logg med posisjon og tidspunkt, dørstenging (tid og sted for åpning og stenging av dører) og hastighet på ulike tidspunkt og posisjoner. I tillegg kan man få data om forhold som energiforbruk. Nyere tog har også i ulikt omfang blitt utstyrt med telling av antallet av- og påstigende passasjerer.

Figur 2 illustrerer økningen over de seneste 10 til 15 årene vedrørende tilgang på punktlighetsdata i jernbanen. Både antallet datakilder og oppløsningen i tilgjengelige data har økt, noe som medfører at jernbanetrafikken kan følges opp på en mye mer detaljert måte nå enn hva som tidligere var tilfelle.



Figur 2. Utvikling av tilgang på punktlighetsdata.

Olsson med flere (2010) viser at det er mulig å kombinere data fra infrastruktursystemer, signalanlegg og eksterne datakilder som værdata. Albayrak (2013) kombinerer data fra signalsystem med GPS-data. Ingen av disse analysene bruker formelt sett Big Data, men de viser at det er mulig å utføre analyser av togtrafikken basert på et bredere utvalg av datakilder enn hva som har vært vanlig hittil, både i evalueringer og i forbedringstiltak.

Tabell 2 er en illustrasjon av indikatorer, kilder og barrierer som kan brukes i evaluering av jernbaneprosjekter. Tabellen tar utgangspunkt i generelle effektmål for store statlige investeringer innenfor samferdsel.

Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
Internett-aktivitet	(opplevd) tid	Omtale av banestrekningen	Internett, sosiale medier	God			
	Aktivitet på toget	Type bruk av internett på toget	Websider som oppsøkes fra lokalt nett	Kan logges. Administrator for wifi-system har tilgang	Viser type internettaktivitet til byggets brukere	Får ikke linkes til den enheten (PC, telefon etc) som brukes	
Bevegelser	Tid	Punktlighet, faktisk reisetid	TIOS, data fra signalsystemene	God	Måler på blokkstrekning	Ikke personopplysning	
	Tid	Punktlighet, faktisk reisetid	GPS	Krever at data lagres, og høyere frekvens enn standard	Måler med høyere oppløsning enn signalanlegg	Ikke personopplysning	
	Tid	Stasjonsopphold	Hastighetslogg, dørstenging	Ikke tradisjon for å levere ut	Måler stasjonsopphold. Kompletterer data fra signalanlegg	Ikke personopplysning	
	Volum	Antall reisende	Passasjetellings-system	Finnes ikke på alle tog. Ikke tradisjon for å levere ut	Måler på-og avstigende reisende	Ikke personopplysning, kun telling	
Fysiske omgivelser	Tid	Vær, temperatur, snøfall	Etablerte værdata	God	Forklaringsfaktor	Ikke personopplysning	Gratis
	Tid	Signalstilling til enhver tid	Signalanlegg	Ikke tradisjon for å levere ut	Kompletterer data om togbevegelser	Ikke personopplysning isolert sett	
	Tid	Tilstand infrastruktur	Logging fra sensorer og posisjonsfølere i infrastrukturen	Ikke tradisjon for å lagre eller levere ut	Kompletterer manuelle registreringer	Ikke personopplysning	
	Tid	Tidsbesparelser	Mobitefondata, fra basestasjon hjemme til basestasjon arbeid	Krever at data gjøres tilgjengelig. Tilnærmet kontinuerlig oppfølging	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggskostnad fremst på sammenstilling og analyse
Kommersiell aktivitet	Volum	Billettsalg	Billettsalg	Bedriftsintern informasjon	Månedskort viser ikke reisevolum, bare omsetning	Må aggregeres	
Interne registre/data	Tid	Feil som påvirker trafikken	Banedata, system med oversikt over infrastrukturen	Intern info hos infrastruktur-forvalter	Viser omfang av bruk og godhet i tekniske løsninger	Ikke personopplysning	
	Driftskostnad	Kostnad for drift- og vedlikehold	Regnskapssystem	Intern info hos infrastruktur-forvalter	Viser kostnadsnivå, livsykluskostnad	Ikke personopplysning	
	Tilstand infrastruktur	Antall feil	Banedata, system med oversikt over infrastrukturen	Intern info hos infrastruktur-forvalter	Viserregistrerte og utbedrede feil	Ikke personopplysning	

Tabell 2. Illustrasjon av mulige indikatorer, datakilder og bruk av nye typer data til evaluering av jernbaneprosjekter. Merk at vurderingene om eksempelvis kostnad og personopplysning er foreløpige. (Se vedlegg C for en større versjon).

5.2 Bygg

Investering i bygg som er underlagt KS-regimet er blant annet såkalte formålsbygg som bygges av Statsbygg. Formålsbyggene er spesielt konstruert for en spesiell type bruk. Det kan være mange ulike typer, som kulturbygg, undervisningsbygg og fengsel. For bygg som tar imot publikum, f.eks. et museum, kan erfaringer fra Big Data i varehandelen være relevant. Varehandelen bruker Big data til å analysere kundeadferd både på aggregert og personnivå. På aggregert nivå identifiserer de hvilke varer som selger best ulike steder og til ulike kundegrupper, både inne i butikken, og mellom butikker og regioner. På personlig nivå kan de skreddersy tilbud til identifiserte kunder basert på kjøpadferden (Davenport 2012). I evaluering av bygg i bruk er det først og fremst prinsippene fra den aggregerte typen av analyser som synes relevante.

Bygg kan evalueres etter flere dimensjoner, inkludert (Vitruvius, 1960) klassiske krav som bygninger skal oppfylle: firmitas (styrke), utilitas (hensiktsmessighet) og venustas (skjønnhet). Driftskostnader (inkludert vedlikehold) er et annet aktuelt aspekt. Evaluering av bygningers hensiktsmessighet og skjønnhet har hittil fremst vært utført med kvalitative metoder (Blakstad, Hansen, Olsson Knudsen 2010).

De fleste store statlige investeringer i bygninger blir bygget og driftet av Statsbygg. Det er aktuelt å skille mellom bygging og drift av selve bygningene, og den virksomhet som utføres i bygningene. Statsbygg er opptatt av det tekniske rundt selve bygget og det driftsmessige (i de tilfeller de også er forvalter, som de som oftest er). Etaten som skal bruke bygget er opptatt av forhold ved bygget som påvirker virksomheten (for eksempel økt produktivitet og samhandling).

For offentlige byggeprosjekter er de generelle effektmålene relatert til brukskvalitet, effektivitet i virksomheten som bedrives i bygget og driftskostnad. Brukskvalitet illustrerer hvordan brukerne oppfatter bygget, og hvordan det brukes. Hva som illustrerer effektiviteten i virksomheten i bygget varierer med type bygg. For sykehus er det ofte eksempelvis knyttet til enhetskostnader for ulike typer av behandling. Ideelt sett skulle man isolere byggets påvirkning på effektiviteten. Bygget kan påvirke effektiviteten på flere måter, inkludert direkte påvirkning på logistikken og den type virksomhet som kan drives i bygget, men også indirekte gjennom trivsel og motivasjon til de som arbeider i bygget. Driftskostnader inkluderer energiforbruk og utgifter til drift, vedlikehold og andre kostnader direkte relatert til selve bygget, eksempelvis ombyggingskostander. Driftskostnadene kan ses på som et mål på effektiviteten i drift av bygget. Indikatorer og datakilder som kan brukes til å følge opp denne type generelle mål er vist i tabell 3.

Det er etatene som utfører behovsanalysene i forkant av etablering av nye bygninger, men virksomhetenes tilbakemeldinger på bruken kunne være viktig også for Statsbygg (for eksempel neste gang de skulle bygge et tilsvarende bygg).

Forhold som er aktuelle å følge opp vedr. bruk av bygg inkluderer:

- Hvor folk er, hvor de samles og møtes, bevegelsesmønster og tidsbruk
- Energi(bruk)/miljøfaktorer
- Komfortsystemer, åpne vindu, skjerme lys, slå på lys, temperatur etc. i bygget

Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
Internett-aktivitet	Brukskvalitet	Opplevelsen av bygget	Omtale av bygningen på internett	Åpent tilgjengelig	Relevant for høyprofilerte bygg som operaen	Ikke personopplysning	
	Effektivitet i virksomhet	Type bruk	Websider som oppsøkes fra lokalt nett	Kan logges. Administrator for wifi-system har tilgang	Viser type internettaktivitet til byggets brukere	Får ikke linkes til den enheten (PC, telefon etc) som brukes	
Bevegelser	Effektivitet i virksomhet, brukskvalitet	Hvor folk er, oppholdstid	Pålogging på lokalt wifi-nett	Kan logges. Administrator for wifi-system har tilgang	Viser utstyr som bruker wifi/internett	Må anonymiseres og/eller aggregeres	Sannsynligvis liten
	Effektivitet i virksomhet, brukskvalitet	Bevegelser, oppholdstid	Adgangskort	Ikke tradisjon for å levere ut	Kun aktuelt for arealer med adgangskontroll	Må anonymiseres og/eller aggregeres	Kostand for aggregering/anonymisering
	Effektivitet i virksomhet, brukskvalitet	Hvor folk er, bevegelser	Videokamera	Krever analyse av video	Viser aktiviteten der det er kamera	Analysen fokuserer på antall, ikke identifisering av enkeltindivider	Sannsynligvis høy
Fysiske omgivelser	Effektivitet i virksomhet, brukskvalitet	Bruk av bygget	Lysbrytere, bevegelsessensorer i rom	Ikke tradisjon for å lagre eller levere ut	Avhenger av type og plassering av sensorer	Ikke personopplysning for felleslokaler	
	Driftskostnad	Energibruk	Energistyringssystemer	Lagres delvis for å kartlegge energiforbruk	Viktig kostnad	Ikke personopplysning for felleslokaler	
	Effektivitet i virksomhet, brukskvalitet	Bruk av bygget	Energistyringssystemer	Lagres delvis	Fokus på energiforbruk, men kan også illustrere bruk	Ikke personopplysning for felleslokaler	
	Brukskvalitet	Inneklima	Klimaanlegg, CO-måler		Del av brukskvalitet	Ikke personopplysning for felleslokaler	
Kommersiell aktivitet	Effektivitet i virksomhet	Antall brukere, type bruk, omsetning	Bruk av betalingskort	Ikke tradisjon for å levere ut	Viktig info for kommersielle lokaler	Personopplysning	
Interne registre/data	Driftskostnad	Vedlikeholds aktivitet	Drifts- og vedlikeholds-systemer	Intern info hos byggets forvalter	Viser omfang av bruk og godhet i tekniske løsninger	Ikke personopplysning	
	Driftskostnad	Kostnad for drift- og vedlikehold	Regnskapssystem	Intern info hos byggets forvalter	Viser kostnadsnivå, livsykluskostnad	Ikke personopplysning	
	Driftskostnad	Omfang av ombygging	Areal register og regnskapssystem	Intern info hos byggets forvalter	Viser tilpasningsdyktighet	Ikke personopplysning	

Tabell 3. Illustrasjon av mulige indikatorer, datakilder og bruk av nye typer data til evaluering av byggeprosjekter. Merk at vurderingene om eksempelvis kostnad og personopplysning er foreløpige. (Se vedlegg C for en større versjon).

Det er utviklet og brukt ulike metoder for vurdering av brukskvalitet i bygninger, eksempelvis USETool (Hansen, Blakstad, Knudsen 2009). Vanlige metoder inkluderer intervjuer og befaringer og spørreskjema til brukerne. Samlet sett kan denne type verktøy gi et godt bilde av hvordan brukerne oppfatter en bygning. Evalueringen blir likevel først og fremst basert på hvordan bygget brukes ved evalueringstidspunktet. I tillegg finnes risikoen for at de brukerne som engasjerer seg i evalueringen ikke er representative, og eksempelvis er de mest eller minst fornøyde brukerne. Big Data åpner for å komplettere kvalitative evalueringsformer. Big Data kan belyse bruk av bygget over lengre tidsrom og dekke en større bredde av brukere. I senere tid har det kommet flere eksempler på evaluering av bygninger i bruk basert på kvantitative metoder med utgangspunkt i Big Data-tilnærminger. Yoshimura med flere (2012) bruker logging av blåtann-enheter (i praksis smarttelefoner) til å beskrive hvordan besøkende beveger seg i museet Louvre i Paris. De beskriver blant annet besøkernes bevegelsesmønster og lengde på besøkene. Rawassizadeh et al. (2011) brukte et kamera for å registrere renheten i et areal. Renheten ble målt ved å

sammenligne fargeintensiteten på et tidspunkt, og bruke fargeintensiteten på en ren flate som referanse. Khanie et al. (2011) brukte utstyr for å følge øyebevegelser. Hensikten var å se på sammenhengen mellom øyebevegelser og opplevd komfort, spesielt relatert til ulike lysforhold.

På samme måte som digitale kart er et hjelpemiddel til å systematisere posisjonsbaserte data på et makronivå (skala i størrelsesorden 1:100 000), kan bygningsinformasjonsmodeller (BIM) brukes til å systematisere data på et mikronivå (skala i størrelsesorden 1:100) til å beskrive bevegelser i en bygning.

Forhold rundt drift og vedlikehold kan evalueres ved bruk av nye data. Det er utført en del kvantitative evaluering av byggekostnader og i senere tid evaluering av energibruk (ZEB 2013). Mulighet for å bruke data fra ulike automasjonsanlegg i et bygg øker når slike anlegg blir vanligere, man får tradisjon for å lagre driftsdata, og man finner måter å bruke disse data på. I tillegg skjer det en utvikling av IT-systemer for å planlegge og følge opp drifts- og vedlikeholdsoppgaver. Eksempelvis innfører Statsbygg nå et nytt system. Dette gir interessante data som nå blir lettere tilgjengelig. Men det er i hvert fall innledningsvis ikke data i så store mengder at det går under definisjonen Big Data, uten at det gjør det til mindre interessant informasjon.

Big Data kan brukes til å kartlegge blant annet hvordan brukerne bruker en bygning, og til å få utvidet og nyansert informasjon om kostnader for forvaltning, drift og vedlikehold.

Med utgangspunkt i sammenstillingen av ulike typer Big Data-kilder anser vi følgende nye type data aktuelle til evaluering av bygninger:

- Ord: Eksempler er hvordan de aktuelle bygningene omtales på internett, Facebook, Twitter etc. (omfang av omtale, tonen: positiv/negativ)
- Stedsdata: hvor mange er i et område, i eller ved bygningen, når på dagen/uken, hvor kommer de fra og hvor drar de. Kan baseres på GPS, mobiltelefoner, adgangskontrollsystemer, videokameraer, eller annet.
- Fysiske omgivelser: Måling av temperatur i bygget, bruk av ulike automasjonsanlegg (lys, klima, energibruk, CO-måler), følere som teller antall passeringer (inn i et areal eksempelvis)
- Adferd: Hva gjør folk, eksempelvis hvilke websider som oppsøkes. Pålogging på datamaskiner kan brukes til å synliggjøre utnyttelsen av kontorarbeidsplasser.
- Økonomisk aktivitet: Registreringer med betalingskort – når, hvordan bruker folk penger?

Tabell 3 utdyper dette.

6 Konklusjon

Med utgangspunkt i de spørsmål som er stilt i prosjektets målsettinger ønsker vi her å oppsummere hovedtrekkene fra studien, og konkludere hvorvidt det er et potensiale for at Big Data kan benyttes i evaluering av store statlige investeringstiltak.

Big Data er et område i rask utvikling. Det publiseres mye, og mange aktører ser muligheter. Svært mye er gjort innenfor privat sektor (varehandel, forretningsanalyse) og da i hovedsak for å utnytte nye kommersielle muligheter, mens offentlig sektor henger noe etter. Det er gjennomført noen få studier i stor skala på temaer som er aktuelle for evaluering av store statlige investeringer, og disse viser at det finnes et potensiale for å bruke nye typer av data (i kategorien Big Data) til evalueringer.

Vår studie har forsøkt å se nærmere på ulike områder som er viktige for potensialet for bruk av Big Data i evaluering; ***personvern, tilgjengelighet, anvendbarhet og relevans, eiendomsrett, kostnader og kompetanse***. Når det gjelder tilgjengeligheten til data vil denne påvirkes av både personvernhensyn, eiendomsrettigheter og kompetanse. Vi konkluderer med at personvernhensyn og eiendomsrettigheter ikke nødvendigvis trenger å være en hindring for bruk og analyse av Big Data. Lovgivningen gjør likevel at man støter på ulike utfordringer som kan gjøre tilgjengeligheten mindre og anvendelsen noe begrenset. Det kan synes som at den teknologiske utviklingen har løpt i fra lovverket. Informasjonsbehovet er stort, og det mangler blant annet internasjonale retningslinjer. Kompetansebehov innen forskning, databehandling, analyse og visualisering er en flaskehals for utviklingen innen området både i Norge og internasjonalt. Dette kan være med på redusere tilgjengeligheten, samt øke kostnadene ved bruk av Big Data. Når det gjelder anvendbarhet og relevans konkluderer vi med at de ulike typene av Big Data trolig har et stort potensiale for å kunne anvendes i evalueringer. Spesielt kan flere ulike datasett som belyser samme fenomen brukes til triangulering og kvalitetssikring av analyser som inngår i evalueringer. Anvendbarheten kan bli noe redusert da data på mange ulike formater ofte trenger stor grad av bearbeiding og tilrettelegging. Dette vil likevel igjen komme an på hvilken teknologi som benyttes. Ved at Big Data skaper nye muligheter til å analysere et fenomen basert på ulike typer av data øker mulighetene for en evaluator til å finne indikatorer som er relevante i forhold til det prosjektet man evaluerer. Både anvendbarhet og relevans utfordres likevel av to forhold. Det ene er at Big Data innebærer en ny måte å forholde seg til informasjon på. Man må tenke nytt når det gjelder bruk av vitenskapelige metoder. Metodiske problemstillinger som representativitet, signifikans, utvalgskriterier, frafall etc. må tilpasses de nye typene av data. Det andre er behovet for data som ligger tilbake i tid, gjerne før tiltaket ble iverksatt. Vil Big Data samlet inn med datidens teknologier kunne sammenliknes med informasjonen man har i dagens situasjon? Disse utfordringene kan reduseres dersom data lagres med høyest mulig oppløsning, og det angis tydelig hvordan dataene er innsamlet og bearbeidet. I tillegg kan det iverksettes tiltak for å sikre at data blir lagret over lengre tidsrom allerede når man starter planleggingen av et prosjekt. Det krever at man kjenner hvilke data som kan være aktuelle, og at man legger til rette for å gjenfinne disse data når det er tid for evaluering etter flere år.

Det synes som at det finnes store muligheter for bruk av nye (store) data i evaluering. Concept kan bli blant de første til å vise muligheter i Big Data i forhold til evaluering generelt, og spesielt relatert til store statlige

investeringer. Vi har funnet flere eksempler på kreativ bruk av Big Data som er relevant for evaluering av store statlige investeringer. Det finnes også interessante data som ikke bokstavelig talt er Big data, men likevel interessant å bruke i evalueringer i større omfang enn hva som synes å være tilfelle hittil. Eksempler på kilder til denne type data er styringssystemer for drift og vedlikehold av offentlig infrastruktur og bygninger.

Når det gjelder de to spesifikke sektorområdene samferdsel og bygg konkluderer vi med at det her finnes muligheter for å fremskaffe et bredere datagrunnlag enn det som typisk utnyttes i dag (både basert på Big Data og andre nye datakilder), som trolig vil kunne både forbedre tidligere måleparametere, samt tilføre nye i evaluering av tiltak. En skjematisk vurdering av aktuelle kilder er gitt i vedlegg C. Tabellen er på ingen måte uttømmende, men kan betraktes som et utgangspunkt for videre arbeid med bruk av Big Data i evaluering av tiltak innenfor samferdsel og bygg.

Vår anbefaling er at det nå blir gjennomført pilotstudier, hvor man forsøker å benytte ulike former for Big Data i en evaluering av et tiltak, for eksempel innen samferdsel og bygg. Concepts rolle kan være å bidra til igangsettelse av slike piloter gjennom å identifisere aktuelle tiltak, aktuelle typer av data, bidra i tolking av data og å sette dem inn i en evalueringssammenheng. Det innebærer at vi trenger bidrag fra aktører med kompetanse på datafangst og Big Data analyser. Dette er et område med meget rask utvikling. For å kunne være blant de første bør derfor pilotene startes så raskt som mulig.

7 Referanser

Albayrak, D. (2013) GPS data utilization for presenting punctuality information in railways. Masteroppgave Norges Teknisk Naturvitenskapelige Universitet.

Alchian, Armen og Harold Demsetz (1972) Production, Information Costs, and Economic Organization. 62 American Economic Review 777-95.

Amazon (2013) Public Data Sets on AWS http://aws.amazon.com/publicdatasets/ Lastet ned 31.7.2013

Arikawa, M., S. Konomi, and K. Ohnishi, Navitime:Supporting pedestrian navigation in the real world, IEEE Pervasive Computing, vol. 6, pp. 21–29, July 2007.

Belassi, W. & Tukel, O.I., (1996); A new framework for determining critical success/failure factors in projects. Management Information Systems, 10(2), 203-255

Berg, P, Andersen, K., Østby, L-E, Lilleby, S., Stryvold, S., Holand, K., et al. (1999). Styring av statlige investeringer. Prosjektet for styring av statlige investeringer. [Management of Governmental Investments]. Finans- og tolldepartementet, Ministry of Finance, Oslo, Norway .

Bertnsen, S. & Sunde, T. (2004) Styring av statlige prosjektporteføljer i staten. Concept rapport nr 1, NTNU

Blakstad, S.H., Hansen, G., Olsson, N., Knudsen, V. (2010). Usability Mapping Tool. CIB World Building Congress. Manchester. May 2010. CIB W111: Usability of workplaces, phase 3. CIB publication, pp 17-29

Calabrese, F., Diao, M., Lorenzo, G., Ferreira, J., and Ratti, C. (2012). Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transportation Research Part C

Cesario,S. (2009). Designing Health Care Environments: Part I. Basic Concepts, Principles, and Issues Related to Evidence-Based Design. The Journal of Continuing Education in Nursing. 40(6), 280-288.

Chomitz, K. M. Big Data for Development: From Information- to Knowledge Societies.

Concept (2013). www.concept.ntnu.no

Computing, vol. 9, no. 1, pp. 13–19, January–March 2010.

Curtis, V. (2007). Dirt, Disgust and Disease: a Natural History of Hygiene. Journal of Epidemiology and Community Health. 61(8), 660–664.

Davenport, T.H. (2012) Enterprise analytics. Optimize performance, process and decisions through Big data. FT Press, Upper Saddle River, NJ

Datatilsynet (2013) http://www.datatilsynet.no/verktoy-skjema/Publikasjoner/Analyser-utredninger/Big Data-er-deg/. Lastet ned 30.7.2013

Difi (2012). Åpne data. Del og skap verdier. Veileder i tilgjengeliggjøring av offentlige data. Direktoratet for forvaltning og IKT. http://data.norge.no/sites/data/files/Veileder-i-tilgjengeliggjoring-av-offentlige-data-V2.pdf

Difi (2013) Åpne offentlige data i Norge http://data.norge.no/. Lastet ned 31.7.2013

Donatelle, R., Snow, C., & Wilcox, A. (1999). Wellness: Choices for Health and Fitness. 2nd ed. Belmont, CA: Wadsworth Publishing Company.

Economist (2013). Don't even think about it. The Economist July 20th 2013, pp 22-23

Facebook (2013) facebook developers, Insights. https://developers.facebook.com/docs/insights/

FAD (2102) Ber etatene slippe fri offentlige data. Nyhet, 31.05.2012. Fornyings, administrasjons- og kirkedepartementet. http://www.regjeringen.no/nb/dep/fad/aktuelt/nyheter/2012/ber-etatene-slippe-fri-offentlige-data--.html?id=683862

Ferris, B., K. Watkins, and A. Borning, Location-awaretools for improving public transit usability, IEEE Pervasive

Forskning.no (2013) Demokrati versjon 2.0. 11. August 2013. http://www.forskning.no/artikler/2013/august/363925

Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, Digital footprinting: Uncovering tourists with usergenerated content, IEEE Pervasive Computing, vol. 7, pp. 36–43, October 2008.

Google (2013), https://support.google.com/analytics/answer/1638635?hl=no

Hansen, G.K., Blakstad, S.H, Knudsen, V. (2009). USEtool, Evaluering av brukskvalitet, METODEHÅNDBOK. SINTEF / NTNU. http://www.metamorfose.ntnu.no/Dokumenter/USEtool

Hilbert, M. (2013). Big Data for Development: From Information- to Knowledge Societies http://ssrn.com/abstract=2205145

Kalil, Tom. "Big Data is a Big Deal". White House. Lastet ned 31.7.2013.

Kitchin, R. (2013). The real-time city? Big data and smart urbanism, GeoJournal. 29. November 2013

Manyika, J. Michael Chui,Brad Brown,Jacques Bughin,Richard Dobbs,Charles Roxburgh,Angela Hung Byers (2011) Big Data: The next frontier for innovation, competition, and productivity.McKinsey Global Institute 2011

MIT (2013) Senseable city lab http://senseable.mit.edu/ lastet ned 30.7.2013

Morgenbladet (2013) Vil revolusjonere samfunnsvitenskapene. Morgenbladet 9. august 2013. http://morgenbladet.no/samfunn/2013/vil_revolusjonere_samfunnsvitenskapene

Narayanan, A. og Shmatikov, V. (2009), De-anonymization Social networks. 2009 30th IEEE Symposium on Security and Privacy. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5207644

Nature Editorial. (2008). Community cleverness required. Nature, 455(7209), 1. doi:10.1038/455001a

Olsson, N.O.E. (2006). Impact analysis of railway projects in a flexibility perspective. Transport Reviews 26:5, 557-569.

Rawassizadeh, R., Khosravipour,, S & Tjoa, A.M. (2011). A Persuasive Approach for Indoor Environment Tidiness. In the 6th International Conference on Persuasive Technology.

Schweitzer, M., Gilpin, L. & Frampton, S. (2004). Healing Spaces: Elements of Environmental Design that make an Impact on Health. The Journal of Alternative and Complementary Medicine. 10(1), 71-83.

Statsbygg (2013). BIM – En kortfattet innføring http://www.statsbygg.no/FoUprosjekter/BIM-Bygningsinformasjonsmodell/BIM-En-kortfattet-innforing/. Lastet ned 30-7-2013

Teknisk Ukeblad (2013). Slipp datasettene fri, de er våre. Kronikk IT. Geir Hansen. Teknisk ukeblad nr 26/22 August 2013, pp 46

Vitruvius, M. (1960). The Ten Books on Architecture. New York: Dover.

Volden, G.H. Samset, K. (2013) Etterevaluering av statlige investeringsprosjekter. Konklusjoner, erfaringer og råd basert på pilotevaluering av fire prosjekter. Concet rapport nr 30. Norges teknisk-Naturvitenskapelige universitet.
http://www.concept.ntnu.no/Publikasjoner/Rapportserie/Nr.%2030_webutgave30_norsk.pdf

Wikipedia (2013) http://sv.wikipedia.org/wiki/Big_data. Lastet ned 30.7.2013

Yoshimura, Y., Girardin, F., Carrascal, J. P., Ratti, C., and Blat, J. (2012).New tools for studying visitor behaviors in museums: a case study at the Louvre. MIT: Senseable city lab http://senseable.mit.edu/ lastet ned 30.7.2013

ZEB (2013) Zero Emission Buildings. http://www.zeb.no/index.php Lastet ned 1.8.2013

Olsson, N., Økland, A., Veiseth,M., Stokland, Ø (2010) Driftsstabilitet på Jernbaneverkets nett – årsaksanalyser 2005-2010. Punktlgihets- og regularitetsutviklingen, granskning av årsaker. SINTEF Teknologi og samfunn. http://www.sintef.no/upload/Konsern/Media/rapport%20jernbane.pdf

8 Vedlegg

Vedlegg A Informanter

Institusjon/bedrift
Atbrox
Datatilsynet
Statsbygg
Byggforsk, NTNU
Institutt for datateknikk og informasjonsvitenskap, NTNU
Kantega
SSB
SINTEF IKT
ITS Norge
FAD
Institutt for Informatikk, Universitetet i Tromsø
Evalueringsforum/DFØ
Statoil
Telenor forskningsavdeling

Vedlegg B Intervjuguide

Intervjuguide Store (og nye) data til evaluering av offentlige investeringer

Om Big data

- Hva forstår du med begrepet "Big data"?
- Innenfor hvilke samfunnsområder kan Big data spille en rolle i Norge i dag og i fremtiden? På hvilken måte?
- Hvilket potensial ligger i Big data for offentlig sektor generelt?
- Hvilket potensial ligger i Big data for evaluering av prosjekter og tiltak? Kjenner du eksempler på bruk i evaluering

Big data i evaluering (både ved utvelgelse av prosjekter (ex-ante) og etter gjennomført prosjekt (ex-post))

- I hvilken grad tror du Big data kan benyttes for å evaluere store statlige investeringstiltak (veiutbygging, kulturhus, tuneller, broer, skolebygg etc.)? På hvilken måte kan de i så fall utnyttes til dette formålet? Kom gjerne med eksempler fra din sektor.
- Hva er de største barrierene/hindringene/utfordringene for utnyttelse av Big data i Norge i dag og i fremtiden?

Noen områder vi er bedt om å undersøke spesielt:
Personvern

- Hva er den største utfordringen med hensyn til personvern når det gjelder utnyttelse og bruk av Big data i dag og i fremtiden?
- Hvilke lover og regler sikrer personvernhensynet i Norge i dag?
- Er det et problem at hensyn til personvern kan komme i konflikt med utnyttelse av data for økt verdiskaping, effektivisering og evt. kommersielle interesser? Hvordan kan man evt. tilrettelegge for å sikre begge hensyn?

Eiendomsrett

- Hva ligger i begrepet "eiendomsrett til data"?
- Hva er den største utfordringen med hensyn til eiendomsrett når det gjelder utnyttelse og bruk av Big data i dag og i fremtiden?
- Hvilke lover og regler sikrer eiendomsrett til data i Norge i dag? Er disse også velegnet for Big data?
- Er det et problem at hensyn til eiendomsrett kan komme i konflikt med utnyttelse av data for økt verdiskaping, effektivisering og evt. kommersielle interesser? Hvordan kan man evt. tilrettelegge for å sikre begge hensyn?

Kostnader

- Hvilke kostnader kommer i tilknytning til bruk av «Big data»?
- Er det noen «skjulte kostander»
- Er kostnader et problem for utnyttelse av data for økt verdiskaping, effektivisering og evt. kommersielle interesser?
- Hvordan kan man evt. tilrettelegge for å redusere kostnadene?

Tilgjengelighet

- Hva er den største utfordringen med hensyn til tilgjengelighet av data?
- Hvilke lover og regler styrer tilgjengelighet av data?
- Finnes det data som kunne brukes til evaluering av offentlige prosjekter, men som har vanskelig tilgjengelighet?
- Hva kan man gjøre for å øke tilgjengeligheten?

Anvendbarhet

- Hvordan er datakvaliteten i store datasett?
- I hvilken grad kan data svekket tillit og i så fall hvorfor?
- Hvordan kan man kontrollere datakvaliteten
- Hvordan kan datakvaliteten evt. bedres??

Relevans

- Hvordan utføres Big data analyser for å sikre relevans i forhold til det man «egentlig vil vite» (eks. top down, finn ut hva man ideelt sett vil ha, og søk etter det, eller bottom up; hvilke data finnens og hva kan man gjøre med det?)
- Hvilke data kan være relevante for å evaluere store statlige investeringer?
- Er det ulike data som er aktuelle ved evaluering av et investeringsprosjekt i tidligfase sammenliknet med en evaluering av samme prosjekt etter gjennomføring
- Kan det finnes data som avdekker eksterne virkninger som tidligere ikke har vært like lett å kartlegge tidligere, eller som ikke var kjent.

Kompetansebehov

- Hvilke kompetansebehov er de mest prekære med hensyn til utnyttelse av Big data i dag og i fremtiden?
- I hvor stor grad er mangel på riktig kompetanse en barriere for utnyttelse av Big data?
- Hva er de viktigste tiltakene man kan gjøre for å sikre tilgang på riktig kompetanse i dag og i fremtiden?

Er det andreviktige forhold vi ikke har spurt om, relatert til bruk av Big Data til evaluering av store statlige investeringer?

Vedlegg C Indikatorer, kilder og barrierer for caseområder.

Sektor	Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
	Internett-aktivitet	(opplevd) tid	Brukertilfredshet, tidsbesparelser	tekstanalyse av medieoppslag	God	Relevant for å komplettere bildet. Utfordringer knyttet til signifikans, endring over tid, tekniske fremskritt.	Ingen, offentlige kilder	Middels. Software til analyse er hyllevare.
Vei		(opplevd) tid	Brukertilfredshet, tidsbesparelser	Analyse av sosiale media, Twittermeldinger, facebook	God	Relevant for å komplettere bildet. Utfordringer knyttet til signifikans, endring over tid, tekniske fremskritt.	Basert på åpent tilgjengelig informasjon	Middels. Software til analyse er hyllevare.
	Bevegelser	Volum	ÅDT, antall biler	Bomstasjon, tellepunkt	Krever at data gjøres tilgjengelig	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	
		Tid	Tidsbesparelser	Bomstasjondata, tid mellom passeringer av to eller flere bommer	Krever at data gjøres tilgjengelig	Måler en viktig effekt, men med få tellepunkter	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggs-kostnad fremst på sammenstilling og analyse
		Tid	Tidsbesparelser	Mobitefondata, fra basestasjon hjemme til basestasjon arbeid	Krever at data gjøres tilgjengelig. Tilnærmet kontinuerlig oppfølging	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggs-kostnad fremst på sammenstilling og analyse
	Fysiske omgivelser	Volum	ÅDT, antall biler	Vegslitasje	Kan måles av Vegvesenet	Kompletterer andre kilder	Ikke personopplysning	Avhenger av Vegvesenet
		Volum	Kjørelengde biler til service og/eller EU-kontroll	Verksteder	Krever at data gjøres tilgjengelig	Viser aggregert kjørelengde. Kan vise kjørelengde for enkeltbiler også	Avhenger av aggregering	Usikkert
		Volum	Luftkvalitet, støy, etc.	Ulike former for målere	Sannsynligvis god, da det typisk er offentlige (kommuner, vegvesen etc som måler)	Måler volum indirekte, kompletterer andre data	Ikke personopplysning	Forhåpentligvis gratis tilgjengelig
	Kommersiell aktivitet	Volum, eksterne effekter	Ringvirkninger	Betalingskort	Finansinstitusjoner	Måler på individnivå eller aggregert	Avhenger av aggregering	Må evt. kjøpes i et kommersielt marked

Sektor	Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
Jernbane	**Internett-aktivitet**	(opplevd) tid	Omtale av banestrekningen	Internett, sosiale medier	God			
		Aktivitet på toget	Type bruk av internett på toget	Websider som oppsøkes fra lokalt nett	Kan logges. Administrator for wifi-system har tilgang	Viser type internettaktivitet til byggets brukere	Får ikke linkes til den enheten (PC, telefon etc) som brukes	
	Bevegelser	Tid	Punktlighet, faktisk reisetid	TIOS, data fra signalsystemene	God	Måler på blokkstrekning	Ikke personopplysning	
		Tid	Punktlighet, faktisk reisetid	GPS	Krever at data lagres, og høyere frekvens enn standard	Måler med høyere oppløsning enn signalanlegg	Ikke personopplysning	
		Tid	Stasjonsopphold	Hastighetslogg, dørstenging	Ikke tradisjon for å levere ut	Måler stasjonsopphold. Kompletterer data fra signalanlegg	Ikke personopplysning	
		Volum	Antall reisende	Passasjetellings-system	Finnes ikke på alle tog. Ikke tradisjon for å levere ut	Måler på-og avstigende reisende	Ikke personopplysning, kun telling	
	Fysiske omgivelser	Tid	Vær, temperatur, snøfall	Etablerte værdata	God	Forklaringsfaktor	Ikke personopplysning	Gratis
		Tid	Signalstilling til enhver tid	Signalanlegg	Ikke tradisjon for å levere ut	Kompletterer data om togbevegelser	Ikke personopplysning isolert sett	
		Tid	Tilstand infrastruktur	Logging fra sensorer og posisjonsfølere i infrastrukturen	Ikke tradisjon for å lagre eller levere ut	Kompletterer manuelle registreringer	Ikke personopplysning	
		Tid	Tidsbesparelser	Mobitefondata, fra basestasjon hjemme til basestasjon arbeid	Krever at data gjøres tilgjengelig. Tilnærmet kontinuerlig oppfølging	Måler en av de viktigste nytte-effektene	Aggregerte og/eller anonymiserte data. Pågår diskusjon om eiendomsrett.	Mye data hentes inn, men brukes lite. Tilleggskostnad fremst på sammenstilling og analyse
	Kommersiell aktivitet	Volum	Billettsalg	Billettsalg	Bedriftsintern informasjon	Månedskort viser ikke reisevolum, bare omsetning	Må aggregeres	
	Interne registre/data	Tid	Feil som påvirker trafikken	Banedata, system med oversikt over infrastrukturen	Intern info hos infrastruktur-forvalter	Viser omfang av bruk og godhet i tekniske løsninger	Ikke personopplysning	
		Driftskostnad	Kostnad for drift- og vedlikehold	Regnskapssystem	Intern info hos infrastruktur-forvalter	Viser kostnadsnivå, livsykluskostnad	Ikke personopplysning	
		Tilstand infrastruktur	Antall feil	Banedata, system med oversikt over infrastrukturen	Intern info hos infrastruktur-forvalter	Viserregistrerte og utbedrede feil	Ikke personopplysning	

Sektor	Kategori	Effekt	Indikator	Eksempel på kilde	Tilgjengelighet	Anvendbarhet og relevans	Personvern og eiendomsrett	Kostnad
Bygg	**Internett-aktivitet**	Brukskvalitet	Opplevelsen av bygget	Omtale av bygningen på internett	Åpent tilgjengelig	Relevant for høyprofilerte bygg som operaen	Ikke personopplysning	
		Effektivitet i virksomhet	Type bruk	Websider som oppsøkes fra lokalt nett	Kan logges. Administrator for wifi-system har tilgang	Viser type internettaktivitet til byggets brukere	Får ikke linkes til den enheten (PC, telefon etc) som brukes	
	Bevegelser	Effektivitet i virksomhet, brukskvalitet	Hvor folk er, oppholdstid	Pålogging på lokalt wifi-nett	Kan logges. Administrator for wifi-system har tilgang	Viser utstyr som bruker wifi/internett	Må anonymiseres og/eller aggregeres	Sannsynligvis liten
		Effektivitet i virksomhet, brukskvalitet	Bevegelser, oppholdstid	Adgangskort	Ikke tradisjon for å levere ut	Kun aktuelt for arealer med adgangskontroll	Må anonymiseres og/eller aggregeres	Kostand for aggregering/anonymisering
		Effektivitet i virksomhet, brukskvalitet	Hvor folk er, bevegelser	Videokamera	Krever analyse av video	Viser aktiviteten der det er kamera	Analysen fokuserer på antall, ikke identifisering av enkeltindivider	Sannsynligvis høy
	Fysiske omgivelser	Effektivitet i virksomhet, brukskvalitet	Bruk av bygget	Lysbrytere, bevegelsessensorer i rom	Ikke tradisjon for å lagre eller levere ut	Avhenger av type og plassering av sensorer	Ikke personopplysning for felleslokaler	
		Driftskostnad	Energibruk	Energistyringssystemer	Lagres delvis for å kartlegge energiforbruk	Viktig kostnad	Ikke personopplysning for felleslokaler	
		Effektivitet i virksomhet, brukskvalitet	Bruk av bygget	Energistyringssystemer	Lagres delvis	Fokus på energiforbruk, men kan også illustrere bruk	Ikke personopplysning for felleslokaler	
		Brukskvalitet	Inneklima	Klimaanlegg, CO-måler		Del av brukskvalitet	Ikke personopplysning for felleslokaler	
	Kommersiell aktivitet	Effektivitet i virksomhet	Antall brukere, type bruk, omsetning	Bruk av betalingskort	Ikke tradisjon for å levere ut	Viktig info for kommersielle lokaler	Personopplysning	
	Interne registre/data	Driftskostnad	Vedlikeholds aktivitet	Drifts- og vedlikeholds-systemer	Intern info hos byggets forvalter	Viser omfang av bruk og godhet i tekniske løsninger	Ikke personopplysning	
		Driftskostnad	Kostnad for drift- og vedlikehold	Regnskapssystem	Intern info hos byggets forvalter	Viser kostnadsnivå, livsykluskostnad	Ikke personopplysning	
		Driftskostnad	Omfang av ombygging	Areal register og regnskapssystem	Intern info hos byggets forvalter	Viser tilpasningsdyktighet	Ikke personopplysning	

Teknologi for et bedre samfunn

www.sintef.no