'Big data' - alternative datakilder og analysemuligheder

Udkast, april 2020

Frederik Hjorth Matt W. Loftis

'Big data' er overalt. Det gælder i dobbelt forstand: takket være drastiske stigninger i computeres hukommelse og regnekraft indeholder næsten alle computere i dag store, ustrukturerede datamængder. Mange af disse data er biprodukter af menneskelig adfærd, som i dag registreres og kvantificeres i historisk uset omfang. Men big data er også overalt i den forstand at begrebet 'big data' og beslægtede begreber er blevet almindeligt kendte og bredt anvendte, og ikke mindst genstand for stor kommerciel interesse. En hyppigt citeret artikel fra *Harvard Business Review* kaldte således "data scientist" for "det 21. århundredes mest sexede job" (Davenport and Patil, 2012).

Alene den begrebslige udbredelse af big data gør det relevant at vide hvad det nærmere dækker over. Men big data er også reelt et væsentligt nybrud i forhold til de data og metoder, politologi og samfundsvidenskab traditionelt har betjent sig af. Big data muliggør analyser af politologiske emner som ville have været umulige med traditionelle metoder, men kræver også nye teknikker og metodiske værktøjer.

Formålet med dette kapitel er at introducere til de datatyper og metoder, begrebet big data dækker over. Først opridser vi begrebets betydning og historie. Dernæst diskuterer vi hvordan en række karakteristika ved big data skaber særlige udfordringer i forhold til at udvikle stærke forskningsdesigns. Herefter præsenterer vi en række specifikke typer af metoder til behandling af big data. Afslutningsvis opridser vi nogle væsentlige etiske problematikker i relation til brugen af big data.

1 Hvad er big data?

I en toneangivende artikel peger Lazer et al. (2009) på big data som kilden til en ny type samfundsvidenskab, "computational social science", med "kapacitet til at indsamle og analysere data med historisk uset bredde, dybde og omfang". Men begrebet big data lever to liv. På den ene side er den populære definition af begrebet forbundet med futuristiske løfter om ny, datadrevet videnskab og teknologi. På den anden side står hvad man kunne kalde den operationelle definition af hvordan store datamænger indsamles, lagres og analyseres af samfundsvidenskabsfolk. For at belyse betydningen af big data for politologi betragter vi først denne anden, operationelle betydning af begrebet inden vi vender tilbage til den første, populære betydning.

En bredt anvendt operationel definition identificerer big data med de såkaldte 'tre V'er': *Volume*, *Variety* og *Velocity*, dvs. omfang, variation og hastighed (Laney, 2001). Big data refererer her til foranderlige og meget store datasæt – inklusive data der er for store til at gemme på en almindelig pc – der

indeholder masser af variation. Datakilder af denne art er ofte interessante for samfundsvidenskab: søgemaskinelogfiler, sociale medieaktiviteter, offentlige registre, mobiltelefonregistre eller endda data gemt ved passiv overvågning udført af digitale enheder i den fysiske verden er alle blevet anvendt til samfundsvidenskabelig forskning (Salganik, 2017). Adgang til disse data kræver partnerskaber med deres ejere – telefonfirmaer, regeringer, teknologiselskaber osv. Dette kan indebære at man skriver software til at få adgang til websteder eller eksterne databaser eller kan involvere mere formaliserede partnerskaber for at dele information sikkert (Einav and Levin, 2014). I afsnittet om "Anskaffelse af data" nedenfor beskriver vi nærmere hvordan det kan finde sted.

Dette peger også på en vigtig forskel mellem big data og traditionelle samfundsvidenskabelige data: traditionelle samfundsvidenskabelige data er typisk indsamlet med samfundsvidenskab som formål. I modsætning hertil omtales big data til tider som "fundne data" eller "digital udstødning" (Harford, 2014). Hvad betyder det? Næsten alle big data er udviklet til *andre formål* end samfundsforskning. Metadata såsom tidsmarkører, antal følgere, eller aktivitetsmål på sociale medier lagres ikke for videnskabens skyld - eller nødvendigvis i overensstemmelse med videnskabelige standarder.

Selv offentlige datakilder kan udvise dette problem. Betragt for eksempel et lille, men illustrativt eksempel: lov nr. 1049 af 11/12/1996. Loven er ganske kort, færre end 20 ord. Det eneste den gør er at ophæve lov om skoleskibsafgift. Civilstyrelsens database med al dansk lovgivning, Retsinformation, angiver adskillige stykker metadata om lov nr. 1049 (jf. `https://www.retsinformation.dk/Forms/R0710.aspx?id=83509`). Metadata angiver f.eks. den lov der ophæves, adskillige relaterede dokumenter, lovens offentliggørelsesdato, og dens ministerområde. Men dette sidste datapunkt er lidt forvirrende. Ministerområdet for lov nr. 1049 er angivet som "Uddannelses- og Forskningsministeriet". Men loven er underskrevet af Mimi Jakobsen, som var erhvervsminister i regeringen Poul Nyrup Rasmussen II. Så hvorfor er loven ikke tilknyttet Erhvervsministeriet? Fejlen opstår fordi Civilstyrelsen opdaterer lovgivning løbende så den afspejler lovgivningens ressortområde *i dag*. Koblingen af lov nr. 1049 til Uddannelses- og Forskningsministeriet er indlysende forkert hvis du skal bruge historiske data om dansk lovgivning. Uheldigvis for politologer er det korrekt hvis du - som Civilstyrelsen - ikke har til formål at bedrive historisk forskning, men i stedet vil organisere gældende dansk lovgivning efter ministerområder. For at gøre ondt værre kan Retsinformation ændre sig yderligere i fremtiden på måder der ikke gavner politologien, alt efter hvad der tjener Civilstyrelsens behov.

Når man analyserer big data der er produceret til et eksisterende privat eller offentligt formål lurer problemer af denne type konstant. Selvom big data kan være enormt værdifulde for samfundsvidenskaben er deres værdi en *utilsigtet bivirkning* af kommerciel aktivitet eller myndighedsudøvelse. Når vi anvender big data i forskningsøjemed er det derfor altid vigtigt at forstå hvorfor og hvordan data er opstået til at begynde med, og tænke igennem hvilke implikationer det har for vores videnskabelige anvendelse. Som Salganik (2017) formulerer det følger både udfordringer og muligheder ved big data af at spørge sig selv hvorfor data blev indsamlet i første omgang.

2 Kilder til big data

Det er ofte en møjsommelig proces at indsamle samfundsvidenskabelige data. Derfor fremhæves det ofte som en fordel ved big data at undersøgelsens subjekter selv genererer data: en forsker kan eksempelvis indsamle millioner af tweets om et politisk emne uden skulle uddele et eneste spørgeskema.

Men selv om data er genereret på forhånd er det ikke ligetil at *anskaffe* sig data. Der er groft sagt tre måder man kan gøre det på.

2.1 Scraping

Den første og mest umiddelbare måde er at udtrække data direkte fra websider, typisk kaldet *scraping*. Scraping udnytter at indholdet på de fleste større websider kommer fra databaser som fremstiller indholdet i websider med en konsistent struktur. Ved at hente kildekoden til disse websider, på samme måde som en webbrowser gør det, kan man udtrække data på en konsistent måde. Hvis man f.eks. besøger websiden for *Lov om ophævelse af lov om skoleskibsafgift* hos Retsinformation finder man i sidens kildekode bl.a. dette:

```
<div class="metadata-summary">
            <span class="kortNavn">LOV nr 1049 af 11/12/1996 Gældende</span><br>
      <div class="ressort">
                Offentliggørelsesdato: 12-12-1996<br>
        Uddannelses- og Forskningsministeriet
            </div>
        </div>
```

Kodestumpen viser at Retsinformations database lagrer lovens navn i feltet `kortNavn` og lovens offentliggørelsesdato og ressortområde i feltet `ressort`. Takket være den stringente kodestruktur er det nemt at gemme disse og andre metadata i et analyserbart format. Og fordi kodestrukturen er ens på tværs af love hos Retsinformation kan man scrape data om tusindvis af andre love med samme lille stykke kode.

Når man indsamler data ved hjælp af scraping tilgår man i princippet data på samme måde som en almindelig internetbruger der benytter sig af en browser. Men fordi scraping gør det muligt at hente kolossale datamængder er det også en kontroversiel praksis. Et illustrativt eksempel på det kommer fra en meget omtalt juridisk strid mellem det sociale netværk LinkedIn og analysefirmaet HiQ. En del af HiQ's forretningsmodel er at analysere arbejdsmarkedet for it-specialister, og HiQ har bl.a. høstet data ved at scrape oplysninger om individuelle it-specialister fra offentlige profiler på LinkedIn. I 2017 sagsøgte LinkedIn HiQ med påstand om at HiQ's scraping-praksis var et brud på amerikansk it-lovgivning. HiQ fik til sidst medhold i at virksomheden kunne scrape data fra offentlige LinkedIn-sider uden tilsagn fra LinkedIn, men sagen illustrerer at scraping ofte finder sted i en juridisk gråzone.

Kodestumpen om *Lov om ophævelse af lov om skoleskibsafgift* i eksemplet ovenfor kommer fra Retsinformation, og det er som hovedregel ikke forbudt at scrape data fra offentlige hjemmesider, så længe man ikke urimeligt belaster udbyderens servere. Man bør dog uanset kilden altid sikre sig tilsagn fra dataudbyderen før man går i gang med at scrape data.

2.2 API'er

En anden måde at hente data på er gennem såkaldte API'er. API står for *Application Programming Interface* og er en slags kontrolleret adgang til data hos en dataudbyder. API'er indebærer altså ikke samme juridiske usikkerheder som scraping, da udbyderen selv stiller data til rådighed og definerer

rammerne herfor. Eksempelvis har mange API'er *rate limits* der sætter grænser for hvor meget data man kan hente ad gangen.

Princippet om at big data ikke er lavet for samfundsforskningens skyld gælder også for API'er. Det egentlige formål for de fleste API'er er at dele data på tværs af kommercielle platforme. For eksempel er der API'er der muliggør at et online-medie kan vise hvilke af ens egne Facebook-venner der har 'liket' en specifik artikel, fordi avisen kan tilgå data om læserens Facebook-netværk gennem Facebooks API. Men mange sociale netværk stiller meget righoldige data til rådighed for forskere gennem API'er. For eksempel bruger Hjorth and Adler-Nissen (2019), som studerer fordelingen af online misinformation, Twitters API til at indsamle data om ca. 13 millioner følgere af ca. 10.000 Twitter-konti. Offentlige myndigheder stiller også i stigende grad data til rådighed gennem API'er. Eksempelvis stiller Folketinget data om medlemmer, forhandlinger og lovarbejde til rådighed gennem en API.

2.3 Datasamarbejder

En tredje måde at få adgang til big data er gennem et egentligt samarbejde med virksomheder der lagrer big data. For eksempel rapporterer Bond et al. (2012) om et eksperiment, hvor samfundsforskere i samarbejde med Facebook randomiserede hvilken type information Facebook-brugere fik om deres venners stemmeadfærd. I kraft af samarbejdet kunne forskerne udføre eksperimentet i en uhørt stor skala: eksperimentet involverede i alt 61 millioner Facebook-brugere.

Studiet af Bond et al. er exceptionelt fordi det kombinerer kvaliteterne ved big data og eksperimentel metode. Mange forskere gør derfor også en stor indsats for at etablere samarbejder med virksomheder og organisationer der kan give dem adgang til data, der ellers ville være utilgængelige. Men samarbejde med virksomheder om big data er ikke uden faldgruber. For det første kræver det ofte et betydeligt bureaukratisk benarbejde at etablere et samarbejde. For det andet, og mere principielt problematisk, er virksomheder og organisationer sjældent interesserede i forskning der stiller dem selv i et dårligt lys. Det kan betyde at nogle typer undersøgelser prioriteres på bekostning af andre, alene fordi de passer bedre til store teknologivirksomheders dagsordener. Eksempelvis konkluderede Bond et al. (2012) at Facebook-kampagnen havde en gunstig effekt på valgdeltagelse. Det er i sagens natur en flatterende konklusion for Facebook. Men det er uklart om forskerne havde haft samme frihedsgrader til at studere de negative konsekvenser af at bruge Facebook.

3 Big data og forskningsdesign

Big data kan anvendes til alle typer empiriske politologiske forskningsspørgsmål, det være sig deskriptive eller forklarende, forudsigende eller kausale spørgsmål. Selvom vi endnu kun er begyndt at se disse anvendelser udfolde sig, har vi i de seneste 15 år fået erfaringer nok til at kunne pege på én vigtig rettesnor når vi anvender big data: forskningsdesign er stadig altafgørende (jf. Toshkov, 2016, s. 13; Clark and Golder, 2015). I afsnittet her diskuterer vi aspekter af forskningsdesign som kræver særlig opmærksomhed når man arbejder med big data. Det drejer sig dels om forskellige typer biases, dels om vigtigheden af at definere sin analyseenhed.

3.1 Velkendt sampling bias

Den måske mest vidtløftige idé om big data kom til udtryk i en nu berygtet artikel i *Wired Magazine* af tidsskriftets stifter, Chris Anderson [anderson08]: idéen om at big data eliminerer stikprøveproblemer fordi '*n*=alt'. Med andre ord, big data gør det muligt at analysere *alle data*, ikke blot en stikprøve. Artiklen har fået stor opmærksomhed og masser af kritik siden den udkom. I begyndelsen af 2020 havde artiklen mere end 2.000 citationer på Google Scholar. For at sætte det i perspektiv har Maurice Duvergers berømte bog *Political Parties*, kilden til den navnkundige Duverger's Lov, samlet lidt mere end 7.000 citationer siden den udkom i 1959.

Som begreb indebærer '*n*=alt' at big data kan aflæses uden forbehold: mønstrene i big data tegner et komplet billede af menneskelig adfærd. Sandheden er imidlertid at vi ikke med big data slipper for at bekymre os om selektionsproblemer. Selektionsproblemer opstår i alle situationer hvor observationer figurerer i data af grunde der hænger systematisk sammen med vores afhængige variabel. Angrist and Pischke (2008) illusterer problemet med et eksempel fra en stor amerikansk folkesundhedsundersøgelse, hvor respondenter i et survey bedes om at angive deres sundhedstilstand på en skala. Ikke overraskende konstaterer forskerne at personer der er indlagt på hospitalet konsekvens udviser et ringere helbred end personer der ikke er indlagt. Skal vi heraf konkludere at hospitaler gør folk mere syge? Svaret er indlysende nej. Det er indlysende, fordi vi i forvejen ved at individer selekterer sig ind i hospitaler netop fordi de er syge til at begynde med. Hvis man ignorerede dette selektionsproblem ville man ende med sampling bias.

Tilsvarende situationer opstår hele tiden i arbejdet med big data. For eksempel beretter Harford (2014) om den amerikanske by Bostons erfaringer med at bruge smartphonedata til at registrere asfalthuller i byens vejnet. En særlig app som indbyggere i Boston kunne installere registrerede bump i vejen ved hjælp af smartphonens indbyggede accelerometer, og kunne på den måde kortlægge byens asfalthuller. Resultatet af projektet var at byen opdagede alle asfalthuller i kvarterer hvor yngre, velstående bilister færdedes – netop den type person som ville være mest tilbøjelig til at eje en smartphone og downloade byens app. Tumasjan et al. (2010) påpeger et lignende problem i tidligere forskning som har forsøgt at forudsige valgresultater med Twitter-data. Twitter-omtale forudsiger ganske rigtig partiopbakning glimrende ved Tysklands forbundsvalg i 2009 – med én slående undtagelse. Det ekstremt online *Piratpartiet* fik en kolossal mængde Twitter-omtale men en meget lille andel af de faktiske stemmer. I begge disse eksempler gælder '*n*=alt' alene for en meget specifik, selvselekteret gruppe af særligt sofistikerede teknologibrugere.

Det har således altid været tilfældet at man risikerer at få de forkerte svar hvis man ikke tager højde for sampling bias. Hvis man bruger big data med en biased stikprøve får man plot et ekstremt præcist estimat af det forkerte svar.

3.2 Særlige problemer ved big data

Big data stiller os ikke blot over for velkendte kilder til bias, men også nye af slagsen. Det måske mest berømte eksempel kommer fra projektet *Google Flu Trends*. Google Flu Trends var et Google-projektet som blev lanceret i 2008 med en ambition om at foudsige regionale influenza-epidemier ved hjælp af finkornede data om Google-søgninger efter typiske inluenza-symptomer (Ginsberg et al., 2009). Til at begynde med demonstrerede GFT en evne til med imponerende nøjagtighed at forudsige influenza-epidemier langt hurtigere end de amerikanske sundhedsmyndigheder kunne. Men over tid blev

GFTs performance ringere. Da projektet blev lukket i 2015 havde GFT i årevis forudsagt langt større influenzaepidemier end dem det amerikanske *Centers for Disease Control and Prevention*, svarende til den danske Sundhedsstyrelse, kunne konstatere (Harford, 2014). Årsagerne til GFTs deroute illusterer både vigtigheden af forskningsdesign generelt og nogle af de specifikke nye udfordringer big data bringer med sig.

Der var flere kilder til GFTs problemer, men det vigtigste lader til at have været Googles egen justering af dets søgemaskine (Lazer et al., 2014). Ligesom alle andre organisationer der producerer eller bearbejder big data ændrer Google fra tid til anden sine algoritmer. I 2011 og 2012 tilføjede Google en funktion til søgemaskinen som foreslår yderligere søgetermer på baggrund af hvilke termer brugeren har indtastet. De nye funktioner foreslog endog mulige diagnoser når brugeren indtastede symptomer på kendte sygdomme. De nye funktioner skabte dermed en feedback-effekt: brugere af Googles søgemaskine blev ansporet til at justere deres søgninger efter symptomer og opsøge information om influenza. Denne informationssøgning forstærkede de signaler GFT brugte til at forudsige influenzaudbrud. Der er flere nuancer i historien, men denne feedback-effekt illustrerer nogle af de centrale udfordringer ved big data. Som Salganik (2017) formulerer det er big data *algoritmisk konfunderet*, *beskidt* og *glidende*.

Algoritmisk konfundering finder sted når computeres adfærd interagerer med menneskelig adfærd i et system og derigennem påvirker den menneskelige adfærd vi interesserer os for. GFT er et eksempel på algoritmisk konfundering, da computerens adfærd – Google-søgemaskinens forslag til søgetermer – interagerede med den menneskelige adfærd GFT brugte til at forudsige influenzaudbrud. Googles forslag til søgetermer ændrede brugernes egen opfattelse af deres symptomer og ansporede dem til en anden søgeadfærd.

Med 'beskidt' refererer Salganik (2017) til det forhold at big data indeholder falske positiver i form af computeradfærd som nemt kan forveksles med menneskelig adfærd. For eksempel har i hvert fald en del af Googles ekstra søgeaktivitet været utilsigtet, og alene forårsaget af søgemaskinens egne forslag. Ejere af en smartphone vil genkende dette problem som en autocorrect-fejl. Et andet, prominent eksempel er tilstedeværelsen af store antal 'bots' på sociale medieplatforme. Bots er brugerkonti som ikke betjenes af mennesker, men i stedet af små stykker software som udfører programmeret adfærd i form af at like, følge eller kommentere andre brugeres opslag. Til trods for Twitter og andre platforms tilbagevendende udrensninger af bots vedbliver der at være et ukendt og formentlig stort antal bots på sociale medier. Enhver analyse af data fra sociale medier må derfor tænke nøje over hvordan man kan rense data for at undgå at forveksle bot-adfærd med menneskelig adfærd.

'Glidning' refererer til de forandringer der fra tid til anden finder sted i måde hvorpå big data indsamles. GFT var offer for glidning i hvordan Google-søgetermer blev indsamlet efter introduktionen af forslag til søgetermer i 2011 og 2012. Helt grundlæggende kom GFTs data før og efter dette skifte fra forskellige populationer – en hvor brugere fik forslag til søgetermer og en hvor de ikke gjorde. I kraft af denne forskel er de to populationer usammenlignelige. Det er her værd at bemærke at GFT var et virksomhedsinternt projekt. Alligevel var ingeniørerne bag GFT ikke opmærksomme på de ændringer i søgemaskinen som deres egne kolleger havde introduceret. Forskere der analyserer big data bør være særligt opmærksomme på hvordan deres datakilder kan være gledet bort fra deres oprindelige karakter.

3.3 Råd til håndtering af biases i big data

Der findes ikke én overordnet kur imod disse biases, men med god forskningsmæssig skik og brug kan du sikre dig, at du fanger dem, før de undergraver validiteten af din undersøgelse. Det første skridt er at gøre dig fortrolig med data. Hvis du indsamler dine data fra en API, bør du læse API'ens dokumentation grundigt. Hvis du scraper dine data fra en webside, bør du sætte dig ind i hvordan websiden vedligeholdes og kurateres – hvem lægger sider op, hvem kan fjerne sider igen, og så videre. Mange organisationer der stiller data til rådighed fortæller om denne slags procedurer, for eksempel gennem blogs eller pressemeddelelser. Brug disse kilder til at lære om forhold i data, som kan udfordre dit forskningsdesign. Og vær ikke bange for at stille opklarende spørgsmål, hvis du er i tvivl. Mange systemadministratorer er glade for at kunne hjælpe en nysgerrig forsker eller studerende.

Dernæst bør du skrive dine antagelser ned og, når det er muligt, teste dem som egentlige hypoteser (Landers et al., 2016). For eksempel kunne GFT have testet for algoritmisk konfundering ved at undersøge om forslag til søgetermer hang systematisk sammen med stigninger i antallet af søgninger efter influenzasymptomer. I en analyse af data fra sociale medier kan man undersøge antagelsen om at data er genereret af mennesker ved for eksempel at tjekke om brugerne ikke skriver indlæg der er ekstremt repetitive, skriver dem med overmenneskelig hastighed, eller på andre måder overskrider rammerne for normal menneskelig adfærd på det sociale medie.

Endelig bør du gøre det til en vane at udføre løbende kontroller af dine data, også når du ikke regner med problemer eller kan formulere en specifik antagelse om dine data. Eftersom big data er så store giver det ikke mening at se direkte på tallene eller teksten i data efter eventuelle problemer. I stedet kan visuelle hjælpemidler være nyttige. Lav scatterplots, histogrammer og andre visualiseringer af dine variable og brug dem til at verificere at mønstrene giver mening. Hvis du støder på et mærkværdigt eller overraskende mønster bør du undersøge det indtil du forstår årsagen. Det vil ofte være glidning eller beskidte data der forårsager det overraskende mønster.

3.4 Analysenheder og måling i big data

Politologiske teorier er altid knyttet til et bestemt analyseniveau. For eksempel foregår en teori om partiadfærd i valgkampe på niveauet 'partier i valgkampe': hvert parti i hvert valgkamp udgør én observation. En teori om individuelle politikeres adfærd i valgkampe omsætter sig naturligt nok i et større antal observationer per valgkamp, siden hver enkelt politiker i den enkelte valgkamp udgør en observation. Det er ret sjældent at politologiske teorier udspiller sig på samme detaljerede niveau som de typisk meget fintmaskede big data. Det er naturligvis principielt muligt at udvikle en teori som kan forklare specifikke tweets, love eller taler, men politologiske teorier udspiller sig i almindelighed på et højere, mere aggregeret niveau.

I næste afsnit opridser vi de mest populære typer af metoder til at arbejde med big data. Som det vil fremgå har disse metoder det til fælles at de i realiteten er målemetoder. Vi betragter metoder til at kategorisere data, afdække mønstre og kondensere store mængder data til enklere, mindre datastrukturer. Det er typisk sådan, at politologer der anvender big data-metoder til deres store datasæt med millioner af tweets eller tusinder af love, nedkoger dem til data der består af eksempelvis tusinder af politiker-måned observationer eller hundreder af regering-år observationer.

Selv når man arbejder med big data er den gængse kvantitative, hypotetisk-deduktive metode i

politologien at konstruere og estimere statistiske modeller med en passende mængde af kontrolvariable. Vi fokuserer derfor i næste afsnit på metodiske valg som er særlige for big data, og som typisk går forud for det mere velkendte kvantitative workflow. Disse metoder drejer sig ofte om at omdanne data til observationer der svarer til den teoretisk definerede analyseenhed. Læsere som er fortrolige med kvalitative metoder vil kende dette analysetrin. For eksempel reduceres lange interviews eller større case-studier ofte til kodninger der opsummerer centrale teoretiske begreber.

Dette peger frem mod en væsentlig pointe om brugen af big data i politologien. Selvom metoder til at behandle big data som sådan i udgangspunktet er computationelle betyder det ikke at forskere partout skal benytte kvantitative metoder til at analysere dem i sidste ende. Deskriptive, kvalitative forskningsdesigns kan med stort udbytte bruges til at analysere outputtet af de metoder vi præsenterer herunder.

4 Behandling af big data

Der er groft sagt to typer udfordringer der gør sig særligt gældende når man arbejder med big data. Der er for det første en mænge *computationelle* udfordringer, dvs. rent tekniske udfordringer med at bearbejde datamængder der er væsentligt større end hvad en almindelig computer typisk skal håndtere. Computationelle udfordringer er hyppige i arbejde med big data. Faktisk afgrænser den oprindelige definition i Laney (2001) netop big data til datamængder der overskrider en almindelig desktop-computers kapacitet. Det er også i praksis en relevant udfordring. Den første frigivelse af data fra Facebooks forskningssamarbejde *Social Science One* bygger eksempelvis på en omtrent en exabyte data, svarende til en milliard gigabytes. Vi dækker ikke håndteringen af computationelle udfordringer nærmere her, men henviser til Varian (2014), som giver en tilgængelig indføring i forskelle tekniske metoder til håndtering af kolossale datamængder.

Vi fokuserer her i stedet på særlige analytiske udfordringer der opstår når man arbejder med big data. Selv hvis vi ser bort fra den rent tekniske håndtering af data er der stadig særpræg ved big data som ofte gør gængse samfundsvidenskabelige metoder utilstrækkelige. Antag for eksempel at vi interesserer os for danske statsministres retorik med afsæt i et datasæt med statsministres nytårstaler mellem 1985 og 2020. Med bare 36 taler er datasættet egentlig ret småt, men ustrukturerede tekstdata af denne type er typiske i big data-analyser, og selv dette lille eksempel illustrerer de særlige analytiske udfordringer ved big data.

4.1 Dimensionalitetsreduktion

Vi kunne for eksempel spørge os selv hvilke ord socialdemokratiske statsministre bruger særligt hyppigt sammenlignet med borgerlige statsministre. En gængs samfundsvidenskabelig tilgang til sådan et problem ville være at estimere en regressionsmodel med statsministerens partitilhørsforhold som afhængig variabel og uafhængige variable der angiver hvert enkelt ords hyppighed i den enkelte tale. Men det kan ikke lade sig gøre. En almindelig regressionsmodel kan kun estimeres hvis antallet af observationer overstiger antallet af uafhængige variable i modellen, og det er ikke tilfældet her. Faktisk har danske statsministre brugt 7.286 forskellige ord i de 36 taler. Datasættet har altså godt og vel 200 gange så mange variable som der er observationer.

Vi har med andre ord at gøre med ekstremt *højdimensionelle* data. Fordi al informationen i data er

spredt ud over et stort antal variable er informationsværdien i hver enkelt variabel paradoksalt nok meget lille. Højdimensionalitet opstår ikke alene i arbejdet med tekstdata, men også med data om sociale netværk eller billeder. Derfor er *dimensionalitetsreduktion* et centralt analytisk mål i næsten alle big data-metoder. Dimensionalitetsreduktion indebærer at man reducerer antallet af variable, som i big data-terminologi ofte kaldes *features*, til et lille antal som indeholder mest mulig relevant information fra den oprindelige, fulde mængde af variable. Man kan tænke på dimensionalitetsreduktion som en måde at udtrække en information om en latent variabel fra et stort antal manifeste variable. Det svarer i princippet til forholdet mellem den teoretiske variabel og indikatorerne i et refleksivt indeks (jf. kapitel 18). Forskellen er blot at antallet af manifeste variable i big data er langt højere.

4.2 Klassifikation og skalering

Det er det konkrete forskningsspørgsmål der afgør hvad der er 'relevant' information i det oprindelige datasæt, og dermed også hvilken metode til dimensionalitetsreduktion der er mest passende. Men et godt udgangspunkt er at gøre sig klart hvilken latent variabel man er interesseret i at måle. Karakteren af denne latente variabel vil være afgørende for hvilken konkret analysemetode man skal tage i brug. En særligt vigtig sondring er om variablen er nominal- eller intervalskaleret.

Hvis den latente variabel er intervalskaleret skal man bruge en metode til *skalering*. Her placerer man altså hver enkel enhed på en numerisk skala der kan fortolkes intervalskaleret. I eksemplet med statsministres nytårstaler kunne man interessere sig for om venstre-højre-ideologi kommer til udtryk i talerne. Her ville en skaleringsmetode placere alle talerne på en venstre-højre-dimension. Metoderne *Wordscores* og *Wordfish*, som præsenteres i kapitel 12, er begge skaleringsmetoder udviklet af politologer.

Hvis den latente variabel er nominalskaleret giver det per definition ikke mening at tilskrive hver enhed en numerisk værdi. Her har vi i stedet at gøre med *klassifikation*, dvs. at hver enhed tilknyttes en teoretisk bestemt kategori. Man kunne for eksempel med afsæt i teorien om emneejerskab interessere sig for hvilke emner statsministre prioriterer i deres nytårstaler. Har har vi at gøre med et klassifikationsproblem, idet målet ikke er at placere alle observationerne på samme kontinuerte skala, men i stedet at knytte hver enkelt observation til en kategori (eller evt. probabilistisk til flere kategorier). I analyser af tekstdata er såkaldte emnemodeller (*topic models*) en meget udbredt metode til at klassificere tekster.

4.3 Superviserede og usuperviserede tilgange

En anden vigtig sondring vedrører graden af 'supervision' i analysen, dvs. i hvilket omfang modellens forudsigelser beror på kendte værdier af den afhængige variabel. Denne sondring figurerer især inden for *maskinlæring*, som er en gren af datalogien som beskæftiger sig med statistiske modeller til forudsigelse. Big data og maskinlæring betragtes tit som parallelle og tæt koblede tendenser i samfundsvidenskaben (se også Bach, Svejgaard and Hjorth, 2019).

På den ene side findes *superviserede tilgange*. I superviseret maskinlæring kender vi en afhængig variabel, typisk betegnet y, som vi ønsker at prædiktere. Med superviserede tilgange ønsker vi at finde frem til den funktion, der bedst kan forudsige y på baggrund af en række variable. Formelt er vores model derfor: $\hat{y} = f(X)$. Her er $\hat{y}$ den forudsagte værdi af y, og målet er at $\hat{y}$ skal være så god en

tilnærmelse af y som muligt. Læringselementet består i, at modellerne selv "lærer" om sammenhængen mellem X og y på baggrund af et datasæt med kendte værdier af y, nogle gange betegnet et *training set*. Et training set kan for eksempel være i form af en stikprøve af data hvor menneskelige kodere har angivet de sande værdier af y. I superviserede tilgange bruger modellen sammenhængene mellem X og y i dette mindre datasæt til at forudsige y i resten af data. Skaleringsmetoden Wordscores, omtalt ovenfor, er et eksempel på superviseret læring idet Wordscores bruger kendte positioner for et lille antal 'referencetekster' til at lære om sammenhængen mellem ord og positioner.

Alternativet er *usuperviserede tilgange*, som adskiller sig fra superviserede tilgange ved at man kun har adgang til X, dvs. de variable der findes i data. Dermed er der ingen kendte værdier af y til at supervisere modellen. I stedet overlades modellen så at sige til på egen hånd at opdage sammenhænge mellem variable. Emnemodeller, omtalt ovenfor, er et eksempel på usuperviseret læring idet de alene klassificerer tekster på baggrund af hvordan ord samvarierer på tværs af tekster, og altså uden at bero på input om teksters 'sande' klassifikationer. Det samme er Wordfish, som modsat Wordscores ikke beror på kendte tekstpositioner.

Fordelen ved usuperviseret læring er åbenlyst at man ikke behøver information om kendte værdier af y, som kan være vanskelig eller ressourcekrævende at skaffe. Ulempen er at man ikke kan vide sig sikker på at modellen indfanger de teoretisk interessante teorier. For eksempel kan en emnemodel opdele teksterne i emner på en anden måde end ens teoretiske forventninger tilsiger. Og under alle omstændigheder kræver det et betydeligt arbejde at fortolke outputtet fra en usuperviseret model. En måde at sammenfatte forskellen mellem superviserede og usuperviserede tilgange er således at med superviserede tilgange ligger fortolkningsarbejdet før modellen estimeres – dvs. i udarbejdelsen af et training set med kendte værdier af y – mens fortolkningsarbejdet med usuperviserede tilgange finder sted efter modellen er estimeret.

5 Etiske problemer ved big data

Med alle de nye muligheder for forskningsdesigns big data-revolutionen har affødt følger også etiske problemer, samfundsvidenskaben ikke før har skullet forholde sig til. Nogle af dem udspringer af nye muligheder for at foretage felteksperimenter uden deltagernes viden eller samtykke. Et berømt og berygtet eksempel er Kramer, Guillory and Hancock (2014), som betjener sig af et felteksperiment til at studere 'emotional contagion', dvs. hvordan følelser spreder sig i sociale netværk. Det gør forfatterne ved i samarbejde med Facebook at manipulere hvilke opslag omkring 700.000 Facebook-brugere ser i deres feed når de åbner Facebook. En tilfældigt udvalgt gruppe blev præsenteret for usædvanligt få positive statusopdateringer fra deres sociale netværk. Brugerne i denne gruppe skrev efterfølgende selv mindre positive statusopdateringer, konsistent med en teoretisk forventning om 'emotional contagion'. (En anden gruppe blev præsenteret for ekstra positive statusopdateringer, og her så man den modsatte effekt).

Eksperimentet af Kramer et al. er kontroversielt fordi forskerne de facto manipulerede følelserne hos flere hundrede tusinder af mennesker uden deres samtykke. Forfatterne henviser selv til at brugere af Facebook når de opretter en profil giver samtykke til bl.a. at indgå i forskningsprojekter. Samtidig er det plausibelt at det havde været ødelæggende for eksperimentets økologiske validitet hvis man havde gjort det tydligt for brugerne at de var en del af et eksperiment, og at en vis grad af hemmeligholdelse

dermed var nødvendig. Alligevel har forfatterne mødt kritik for eksempelvis ikke at have debriefet brugerne om eksperimentet *efter* det fandt sted.

Kritikken af dette og tilsvarende eksperimenter på online platforme kompliceres af at eksperimenter allerede finder sted i stor skala på alle de mest populære platforme. Virksomheder som Google, Facebook og Twitter udruller løbende eksperimenter for at lære om effekter af forskellige justeringer, og alle disse små eksperimenter finder sted uden brugernes viden eller samtykke. Man kan som modargument hævde at samfundsvidenskab bør operere efter strengere etiske krav end tech-giganter. Uanset ens position i denne debat er det en kendsgerning at samfundsvidenskaben i big data-æraen har mistet sit traditionelle monopol på eksperimentelle studier af adfærd. I stedet er studier såsom Kramer, Guillory and Hancock (2014) blot en forsvindende lille del af de eksperimenter vi alle sammen indgår i samme øjeblik vi går online.

En anden etisk overvejelse i relation til big data drejer sig om transparens. Som vi så ovenfor bruger forskellige big data-metoder information om sammenhænge mellem features – og evt. mellem features og kendte værdier – til at skalere eller klassificere observationer. Superviserede metoder lærer om relationer mellem features og kendte kategorier, mens usuperviserede metoder bruger multidimensionelle sammenhænge i data til at estimere skalaer og kategorier. Det er vigtigt at erindre at de sammenhænge i data, disse metoder lærer om, ikke er neutrale. Det skyldes at de tilgrundliggende data i sig selv ikke er neutrale. Data udspringer af den virkelige verden, og alle de biases der følger heraf, i form af både biases i grundlæggende menneskelig kognition og fordomme mod sociale grupper, kommer så at sige med i købet.

Implikationen er at ukritisk læring fra data kan risikere at reproducere disse biases, et problem man kalder *algoritmisk bias*. For eksempel beretter O'Neill (2016) om hvordan amerikanske kreditkortfirmaer markedsfører deres produkter online ud fra en såkalt *e-score*. En e-score udregnes på baggrund af big data, som bl.a. inkluderer en potentiel kundes browserhistorik, geografiske position, og andre datapunkter. Det har den konsekvens at internetbrugere fra fattige egne af landet tildeles en lavere e-score. Og som konsekvens af denne lave e-score præsenteres disse brugere for kreditkortprodukter med højere renteomkostninger. På den måde reproducerer algoritmen økonomisk ulighed og kan endda reproducere racemæssig ulighed i samfund hvor racetilhørsforhold og økonomisk status i forvejen er tæt forbundne.

Enhver algoritme som lærer om mønstre på baggrund af faktisk menneskelig adfærd vil opsamle sammenhænge af denne type – sammenhænge som i sig selv er produkter af historiske uligheder med afsæt i eksempelvis race, køn eller etnicitet. Algoritmisk bias er i særdeleshed relevant i studiet af politik og offentlig forvaltning. Staten foretager hver dag indgribende beslutninger med store konsekvenser for borgerne. Det gælder for eksempel når socialrådgivere skal afgøre om et barn skal anbringes uden for hjemmet, eller når jobcentre skal hjælpe ledige tilbage på arbejdsmarkedet. Hvis man trænede en klassifikationsmodel til disse problemer på historiske data ville en sådan model arve alle de biases og uligheder der har præget historiske afgørelser på området.

Her er transparens en afgørende værdi. Hvad end man betjener sig af big data til forskning, kommerciel virksomhed eller myndighedsudøvelse er det afgørende at granske sine modeller for algoritmisk bias. Udfordringen er her at mange modeller i big data er så komplicerede at de er vanskelige eller endog umulige at gennemskue. Det står i modsætning til den relativt enkle logik i eksempelvis lineær regression (Castelvecchi, 2016). Derfor er det op til undersøgeren at granske modellen efter bias ved at

efterprøve modellens forudsigelser under forskellige kombinationer af inputs. Algoritmisk transparens nyder stor opmærksomhed som forskningsområde i datalogien. Men der findes ingen teknik der kan gøre det ud for omhyggelig efterprøvning af modellens forudsigelser og en grundlæggende fortrolighed med data.

References

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bach, Alexander, Jesper Svejgaard and Frederik Hjorth. 2019. "Maskinlæring som politologisk værktøj." *Politica* 51(2).

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295–298.

Castelvecchi, Davide. 2016. "Can we open the black box of AI?" *Nature* 538.

Clark, William Roberts and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?: Introduction." *PS: Political Science & Politics* 48(1):65–70.

Davenport, Thomas H and DJ Patil. 2012. "Data scientist." *Harvard business review* 90(5):70–76.

Einav, Liran and Jonathan Levin. 2014. "Economics in the age of big data." *Science* 346(6210).

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski and Larry Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457(7232):1012–1014.

Harford, Tim. 2014. "Big data: A big mistake?" *Significance* pp. 14–19.

Hjorth, Frederik and Rebecca Adler-Nissen. 2019. "Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences." *Journal of Communication* 69(2):168–192.

Kramer, Adam DI, Jamie E Guillory and Jeffrey T Hancock. 2014. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences* 111(24):8788–8790.

Landers, Richard N, Robert C Brusso, Katelyn J Cavanaugh and Andrew B Collmus. 2016. "A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research." *Psychological methods* 21(4):475.

Laney, Doug. 2001. "3-D Data Management: Controlling Data Volume, Velocity and Variety." *Application Delivery Strategies* .

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915):721–723.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176):1203–1205.
URL: *https://science.sciencemag.org/content/343/6176/1203*

O'Neill, Cathy. 2016. "Weapons of Math Destruction." *Discovery* .

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Toshkov, Dimiter. 2016. *Research design in political science*. Macmillan International Higher Education.

Tumasjan, Andranik, Timm O Sprenger, Philipp G Sandner and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.

Varian, Hal R. 2014. "Big data: New tricks for econometrics." *Journal of Economic Perspectives* 28(2):3–28.