

‘Big data’ og politologisk datavidenskab

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 tekniske værktøjer 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 datamængder 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 has applications in political science in empirical studies of all types, from description to explanation, prediction, and causal studies. Although these applications are only beginning across the social sciences, the past 15 years have provided enough experience that we can already point to one strong finding that can always guide us when we apply big data in our work: research design still matters (see Toshkov, 2016, p. 13; Clark and Golder, 2015). Here we discuss aspects of research design that deserve special attention when working with big data, namely, several biases and the definition of the unit of analysis.

3.1 Familiar sampling bias

Perhaps the headiest idea about big data was expressed best in a, now infamous, article in *Wired* magazine by Chris Anderson (2008) – the idea that big data eliminates sampling problems because $n=\text{all}$. That is, one can analyze *all of the data*. The article has met with a good deal of attention and pushback in the years since it was published. In early 2020, it boasts more than 2.000 citations on Google Scholar. To put that in perspective, Maurice Duverger’s famous book *Political Parties*, the origin of the eponymous Duverger’s Law, was published in 1959 and counts just over 7.000 citations.

As a concept, $n=\text{all}$ implies that big data can be taken at face value. Its patterns reveal the complete picture of human behavior. Unfortunately, the selection problem is still with us. It arises whenever observations enter your data for reasons that systematically relate to the outcome variable. Angrist and Pischke (2008) illustrate the problem with an example from the United States’ National Health Interview Survey, in which individuals are asked to rate their health on a scale. Perhaps unsurprisingly, they find that individuals in hospital consistently rate their health as worse than those out of hospital. Then they pose an interesting question: should we conclude that hospitals make people sicker? The answer is, of course, no. We know that because we already understand that individuals in hospital *selected* into being there precisely because they were sick. Failing to recognize that would result in sampling bias.

Similar situations regularly arise with big data. For example, Harford (2014) recounts the story of Boston’s experiment with using cell phone tracking to identify pot holes in the city’s streets that need repair. The result was that the city discovered every pothole in neighborhoods frequented by young, affluent drivers—the profile of the type of person who owned a smart phone and downloaded the city’s app. Uncovering a similar problem, Tumasjan et al. (2010) confirm a finding in previous research that Twitter mentions of German parties correlated strongly with their vote share in the 2009 parliamentary elections, with one major exception. The extremely online *Pirate Party* garnered a huge number of mentions on Twitter while receiving a tiny fraction of the actual vote. In both of these cases, if $n=\text{all}$ then it equals “all” of a specific, self-selected group of technology users.

Ignoring sampling bias has always meant that researchers risked finding the wrong answer. Using big data with biased samples simply means now we get extremely precise estimates of the wrong answer.

3.2 Special headaches for big data

Big data confronts us with new sources of bias and confounding. Perhaps the most famous example for social scientists is that of Google Flu Trends (GFT). GFT was a Google project launched in 2008 that predicted regional flu epidemics from fine-grained data on users' Google queries about likely flu symptoms (Ginsberg et al., 2009). After initially receiving attention for its impressive accuracy, GFT's predictive performance began to diminish over time until—by the time the project was shuttered in 2015—for years it had produced inaccurately high forecasts sometimes as great as double the number of flu cases reported by the U.S. Centers for Disease Control and Prevention (Harford, 2014). The reasons for GFT's collapse are instructive for understanding both why research design still matters and what new headaches come with big data.

Several things contributed to GFT's problems, but the most noteworthy seems to have arisen after Google adjusted its main search service (see Lazer et al., 2014). Like any organization that produces or processes big data, Google changes its services or its algorithms from time to time. In 2011 and 2012, Google added functions to its search service that suggested additional search terms to users, based on their original search terms. It even suggested possible diagnoses for search terms that involved disease symptoms. The upshot was an apparent feedback loop: users of Google search were nudged to refine their searches for symptoms and to seek out information on the flu, intensifying the signals that GFT relied on to predict flu outbreaks. There is more to GFT's story, of course, but this part of it points out some special features of big data researchers must consider to avoid sampling bias. As Salganik (2017) puts it, big data is *algorithmically confounded*, *dirty*, and *drifting*.

Algorithmic confounding happens when computer behavior interacts with human behavior in a system, altering the human behavior we want to study. GFT is an example, since computer behavior—the Google search algorithm's recommendations—interacted with the human behavior that GFT relied on to predict flu outbreaks (searches for flu symptoms). Google's search suggestions altered users' perceptions of their symptoms and led them to different search behavior.

By “dirty,” Salganik (2017) means that big data also contains false positives in the form of computer behavior that is mistaken for human behavior. For example, at least a portion of Google's additional search traffic around flu symptoms may have been purely accidental, caused only by the search engine's suggestions. Smart phone users will recognize this problem as an autocorrect fail. An example that occasionally makes headlines would be the presence of large numbers of “bots” on social media platforms. Bots are accounts that are not operated by human users, but rather are simply software programs executing automated behaviors in the form of liking, following, or commenting. Despite Twitter and other platforms' occasional purges of bot accounts, an unspecified and probably high number of bots persist. Any analysis of social media data must think carefully about how to clean the data to avoid analyzing bot behavior as if it were human behavior.

Drift refers to the occasional changes that occur in how big data are collected, a problem we foreshadowed in the first section. GFT was a victim of the drift in Google's search data that occurred after the introduction of search suggestions in 2011. Fundamentally, GFT's data before and after 2011 came from different populations—one in which users were treated with nudges and the other without that treatment. Therefore, these data were impossible to compare. Take note. GFT was an internal project, and yet the GFT team missed changes that their Google search colleagues introduced. Researchers using big data must stay aware of how their data may have drifted.

3.3 Overcoming biases in big data

These biases have no general cure, but adopting good practices can ensure you catch many problems before they become threats to the validity of your research design. First, get to know your data. If you collect your data from an API, read its documentation carefully. If you take it from a website, then understand everything you can about how the data on the site are maintained and curated—who puts it up, who can remove it, how it is updated, etc. Many organizations that produce and provide big data communicate about it, for example on company blogs or through press releases. Pay attention to these sources of information for things that can affect your research design. When in doubt, ask clarifying questions. Many system administrators or database managers are happy to help courteous researchers.

Second, try to write down your assumptions explicitly and, whenever you can, directly test them like hypotheses (see, Landers et al., 2016). For example, GFT could have tested for algorithmic confounding by examining whether feedback from the search engine to the user was associated with increases in symptom-related search terms. Social media analyses should test the assumption that their data are generated by humans by checking, for example, that individuals in the data do not post too repetitively, do not post inhumanly fast, and exhibit use patterns that are within some range of normal behavior for the platform.

Finally, develop a habit of running sanity checks, even when you do not anticipate problems and cannot formulate an explicit assumption. Since big data is, well, big, it is not realistic to look at the numbers or text in your data and expect to see anything useful. We do this by drawing pictures. Make scatterplots, histograms, and other plots of your variables and be sure the patterns make sense. When you discover a strange pattern, investigate it until you understand it. You may very well uncover hidden drift or the tell-tale patterns of dirty data caused by bots or other automated interference.

3.4 Levels of analysis and measurements on big data

Theories in empirical political science come with a *level of analysis*. For example, a theory about party behavior during election campaigns has a party-election level of analysis: each party at each election is one observation. A theory about politicians during election campaigns gives a necessarily larger number of observations per election, since each individual politician during a campaign is one observation. It is relatively rare in political science to express our theories at the level of granularity in most big data sets. It is, of course, entirely possible to develop a theory to explain individual tweets or individual laws, etc., but political science generally expresses its theories at a higher level (less granular) level of analysis.

In the next section, we introduce several popular and useful methods for working with big data. As we will see, these methods specifically associated with big data have in common that they are actually measurement tools. We will look at methods for labeling data, discovering patterns, and condensing large amounts of data into simpler, smaller data. It is typical for political scientists applying big data that their “big” data set of millions of tweets or thousands of laws condenses into a data set of, for example, thousands of politician-month observations or hundreds of government-year observations.

Even when working with big data sets, the workhorse method of observational, quantitative hypothesis-testing in political science remains to build and estimate appropriate statistical models with appropriate control variables. Thus, we will focus on the methodological choices that distinguish analyses using big data from the more familiar social science workflow of building a purpose-built data set arranged at the correct level of analysis. Readers who are familiar with qualitative methods will be more familiar with this pattern. For example, lengthy interviews or wide-ranging case studies are often condensed into a summary coding of theoretical concepts.

This raises a point worth emphasizing about big data methods in political science. Although methods for working with big data have to be computational, they do not necessarily mean that researchers must apply quantitative methodology in their overall research design. Descriptive and qualitative research designs can equally well apply the methods we discuss below.

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 mængde *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 Dimensionalitätsreduktion

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.

MAYBE VISUALIZE DFM HERE

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 *dimensionalitätsreduktion* et centralt analytisk mål i næsten alle big data-metoder. Dimensionalitätsreduktion 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 dimensionalitätsreduktion som en måde at udtrække en information om en latent variabel fra et stort antal manifesterede variable. Det svarer i princippet til forholdet mellem den teoretiske variabel og indikatorerne i et reflektivt indeks (jf. kapitel XX). Forskellen er blot at antallet af manifesterede 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 dimensionalitätsreduktion 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 variabelen 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 XX, 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. I behandling af tekstdata **TODO: finish this**

TODO

4.3 Superviserede og usuperviserede tilgange

TODO

5 Etiske problemer ved big data

emotional contagion example: Kramer, Guillory and Hancock (2014)

problem: lack of informed consent

complication: under surveillance capitalism, all citizens are subject to constant experimentation w/o consent

A final ethical concern to consider when applying big data relates to transparency and human bias. As we saw above, methods for making measurements from big data take their results directly from the data. Supervised classification learns associations between features and the target outcome, while unsupervised methods mine multidimensional associations in the data to return scales, clusters, etc. It is valuable to keep in mind that simply learning the associations in data is not neutral because data are not neutral. They are produced in the real world, complete with all of its biases: from the quirks of human perception to systematic racial prejudice or social inequality.

Uncritically learning associations in data can mean doubling down on negative aspects of the real world. For example, O'Neill (2016) tells the story of how credit card companies in the U.S. market their products online to consumers, based on something called an *e-score*. The e-score is calculated from big data, including a potential customer's browsing history, location, and other data points. As a result, for example, people browsing the Internet from poorer areas of the country, with higher loan default rates, would generally receive a worse e-score. Lower e-scores mean the person sees advertisements for credit cards with less favorable terms, reflecting their apparent risk for the card issuer. This pattern contributes to deepening economic inequality and even racial disparities in wealth when neighborhoods are strongly racially segregated.

Any algorithm that learns from patterns in real-world data on human behavior can and will pick up associations of this nature – legacies of inequality, racial and gender prejudice, etc. This is particularly relevant for the study of political science and public administration. The state makes life-altering decisions each day, for example, social workers help decide whether to remove children from potentially abusive homes into protective state custody or and case workers help match unemployed job-seekers to employment opportunities. If one were to apply a classification model, trained on historical data, to assist these processes, that model would inherit all the systematic biases or inequalities that may have characterized that system.

This is where transparency is crucial. Whether using big data for research or applying it in business or government, when issues like fairness or prejudice matter, it is crucial that we examine our models for these types of biases. This means slowing down, examining a model's results for bias, and being transparent about what we find before rushing to apply it. Many methods applied to big data are difficult or impossible to directly examine unlike, for instance, the coefficients of a regression model (Castelvecchi, 2016). Therefore, it is up to the analyst to probe the model for evidence of bias by studying what the model predicts or what measurements it returns under a variety of circumstances. Algorithmic transparency is an active research field in computer science, but unfortunately right now there is no methodological trick that can substitute for carefully and systematically examining our models' performance.

References

- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired* .
- Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- 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, Dimitar. 2016. *Research design in political science*. Macmillan International Higher Education.
- Tumasjan, Andranik, Timm O Sprenger, Philipp G Sandner and Isabell M Welp. 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.