

‘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.

cites: Mullainathan and Spiess (2017), Varian (2014)

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.

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 data 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* 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.

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 det 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 rækkevidden 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.

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 m.fl. er exceptionelt fordi det kombinerer kvalitetene 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).

Familiar sampling bias

Perhaps the most heady idea about big data was expressed best in the, now infamous, article in *Wired* magazine by Chris Anderson (2008) – the idea that big data eliminates sampling problems because $n=all$. 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 book *Political Parties*, the origin of political science's only finding that rises to the level of a law, was published in 1959 and counts just over 7.000 citations.

The concept of “*n*-all” implies that big data can be taken at face value. Its patterns reveal a complete picture of human behavior. Unfortunately, the selection problem is still with us. It arises whenever observations enter your data for systematic reasons. Angrist and Pischke (2008) illustrate the problem with an example from the National Health Interview Survey in the US, in which individuals are asked to rate their health on a scale. Perhaps surprisingly, they find that individuals in hospitals systematically rate their health as worse than those out of hospital. Is the obvious conclusion that hospitals make people sicker? Of course not, individuals in hospital *selected* into being in the hospital because they were sicker.

Similar situations often arise with big data. For example, Harford (2014) recounts the story of Boston’s experiment with using cell phone tracking to identify potholes in the streets that need repair. The result was that the city discovered every pothole in neighborhoods frequented by young, affluent drivers—they type of person who owned a smart phone and downloaded the city’s app. Likewise, Tumasjan et al. (2010) confirm the 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.

Special headaches for big data

Big data also 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 queries to Google about their symptoms (Ginsberg et al., 2009). After initially receiving attention for its impressive accuracy, GFT’s predictive performance began to diminish systematically over time until—by the time the project was shuttered in 2015—for years it had produced inaccurately high estimates sometimes as great as double the data reported by US 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). Google, like many organizations that produce and process big data, is constantly updating its service. In 2011 and 2012, Google added functions that suggested additional search terms to users (based on their initial search terms) and even suggested possible diagnoses for search terms that involved symptoms. The upshot was a 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 highlights some important principles for researchers to keep in mind.

Big data in political science research designs: - addressing confounding - it’s dirty (re: Salganik) – human behavior is mixed together with actions taken by bots/automated systems - drifting – Needs of *actual* system maintainers / users may be completely orthogonal to needs of political science research (or even opposed, consider FB) - algorithmically confounded – large-scale systems have robot nannies.

These include everything from spell checkers to YouTube's recommendation algorithm. The observed behavior we find in big data is a consequence of the (generally) unobservable interaction between humans and these algorithms.

3.1 measurement stuff

- role of big data in the research design
 - uncommon to directly analyze
 - more often: measurement

4 Behandling af big data

Big data præsenterer både computationelle og analytiske udfordringer - vi fokuserer på analytiske

Fællestræk: højdimensionalitet

Fælles for metoder: dimensionalitetsreduktion

Klassifikation ctr. skalering

usuperviserede tilgange

superviserede tilgange

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

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.
- 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.
- 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>
- Mullainathan, Sendhil and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2):87–106.
- 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.