'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 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.

Even government data sources present this problem, sometimes in confusing ways. Consider a minor example: lov nr 1049 af 11/12/1996. The law is short, fewer than 20 words. The only thing it does is ophæve lov om skoleskibsafgift. Civilstyrelsens database of all Danish law, *retsinformation.dk*, lists several pieces of metadata about lov nr 1049.[1] For example, the metadata capture the law it repeals, several related documents, its publication date, and ministerområdet. This last one is confusing, however. Ministerområdet for lov nr 1049 is currently listed as Uddannelses- og Forskningsministeriet. However, the law was signed by Mimi Jakobsen. If you do some further research, you will find that Mimi Jakobsen var Erhvervsminister i Regeringen Poul Nyrup Rasmussen II.[2] Why, then, is this law not connected to Erhvervsministeriet? This confusion is caused because Civilstyrelsen opdaterer ministerområder løbende based on *today's* legal framework and ministries. Connecting lov nr 1049 to Uddannelses- og Forskningsministeriet is obviously wrong if you want to build a historical data set of Danish law. Unfortunately for political scientists, it is correct if—like Civilstyrelsen—your goal is not historical research, but instead to organize the legislation currently in force in Denmark. To make matters worse for researchers, retsinformation.dk will continue to evolve in ways that may or may not benefit political science research, depending on Civilstyrelsens future needs and goals.

When using big data built for some official or private purpose, problems of this nature are always lurking. Although big data can be extremely useful for social science, their scientific value is an *unintended byproduct* (i.e. exhaust) of business or government activity. When applying big data to social science research we must always understand why and how the data were assembled and probe the implications of the data's purpose for our scientific applications.[3]

[1] See: https://www.retsinformation.dk/Forms/R0710.aspx?id=83509

[2] https://www.regeringen.dk/regeringer-siden-1848/regeringen-poul-nyrup-rasmussen-ii/

[3] As Salganik (2017) puts it, the challenges and opportunities created by big data follow from asking why the data were collected.

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 efter samme princip.

3 Big data og forskningsdesign

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. - Standard sampling issues (nonrepresentative, systematic sampling bias) - its role - uncommon to directly analyze - more often: measurement

4 Behandling af big data

4.1 Anskaffelse af data

- Getting data: scraping, APIs, text databases

4.2 Usuperviserede tilgange

4.3 Superviserede tilgange

4.4 Tekst som data

- Pattern discovery (dimensionality reduction – a la argument in Lowe 2013WP)
  - Generic data: Clustering, IRT, etc.
  - Fundamental unity of goals and approach
  - Variety in methods results from variation in:
    - Assumptions re: underlying model/geometry of latent space
    - Related to above: something like, "format" of output
    - Structure/nature of input data
    - Amount of domain expertise applied to structure results
    - Assumptions about what is correlated with what
    - Level of computational intensity
- Text: Topic modeling, text scaling, dictionaries
- Classification / prediction
- Explanatory modeling

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

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).

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

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.

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.

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