‘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 beytdning 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)

0.1 Hvad er big data?

Big data has been credited with enabling an emerging field of computational social science with “the capacity to collect and analyze data with an unprecedented breadth and depth and scale” (Lazer et al., 2009). Yet, the concept of big data lives two lives. On the one hand the popular definition is associated with exciting and futuristic promises of data-driven science and technology. On the other hand is what we might call the operational definition of how massive data sets are collected, stored, and used currently by social scientists. To unpack the role of big data in political science research, we first consider the second aspect of the concept and then circle back to examine what to believe about the first aspect.

One widely familiar operational definition identifies big data with the so-called three Vs: Volume, Variety, and Velocity (Laney, 2001). Big data, here, refers to fast-evolving and very large data sets–including even data too large to store on a desktop computer–that contain lots of variation. Data sources of this nature are often interesting to social science: search engine logs, social media activity, government administrative records, mobile telephone records, or even data stored by passive monitoring conducted by digital devices in the physical world have been mobilized for social science research (Salganik, 2017). Accessing these data requires partnerships with their owners–phone companies, governments, technology companies, etc. This can mean writing software to access websites or remote databases or can involve more formalized partnerships to share information securely (Einav and Levin, 2014).

The hurdle of accessing big data underscores a basic difference between it and traditional social science data: traditional social science data were collected for the purpose of doing social science. Social scientists must always consider the sources and the nature of our data, and big data has sometimes been referred to as "found data" or "digital exhaust" (Harford, 2014). What does this mean? Virtually all big data are purpose-built for goals *other than* social research. Metadata like timestamps, follower counts, or activity frequencies on social media web sites are not stored for science or, necessarily, according to scientific standards. Although they may be useful, their scientific value is an unintended byproduct (i.e. exhaust) of business or government activity. As such, when applying big data to social science research we must always probe the implications of the data's purpose for our scientific applications.[1]

- How/why big data became what it is in the zeitgeist
  - Classic advantages: Big / always on
  - Classic disadvantages: metered/restricted access + expertise barrier

0.2 Big data og forskningsdesign

Systems that collect big data are purpose-built, and the purpose is never political science research. This holds true even for the most research-friendly data sources.

- Big data typically *not designed for research*, i.e.
  - 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)

0.3 Behandling af big data

- Pattern discovery (dimensionality reduction – a la argument in Lowe 2013WP)

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

    - Generic data: Clustering, IRT, etc.
    - Text: Topic modeling, text scaling, dictionaries
    - 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

- Classification / prediction

- Explanatory modeling

0.4 Etiske problemer ved big data

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.

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.