‘Big data’ og politologisk datavidenskab

Udkast, januar 2020

Frederik Hjorth
Matt W. Loftis

Draft motivation:

Training, job market demands, sexy research, etc. increasingly value the ability to work with big data. It’s not always clear what one means by big data or how one works with it, but there is a consensus that doing so requires skills. We agree, and we present some of these skills here, along with a description of what big data means and in what ways we work with it.

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

What is big data?

- Operational definition
  - e.g. “Volume, Variety, and Velocity”; Harford (re: “found data”)
  - ‘digital exhaust’ / metadata vs. purpose-built systems
- How/why big data became what it is in the zeitgeist
  - Classic advantages: Big / always on
  - Classic disadvantages: metered/restricted access + expertise barrier

First skill of working w/big data: *Research Design*

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)

What to do with it:

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

  - 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

References

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

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