

Big Data and Security

Moderní technologie a bezpečnost (BSSn4411)

Modern technologies and conflict (CDSn4003)

Jan Kleiner

2.11.2020


M U N I
F S S

Presentation outline



- Big Data – theoretical and methodological prism.
- Legitimate ways of use.
- Problematic ways of use.

- **„Big Data refers to datasets, whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse.“**

- 
- **Definition intentionally subjective and moving.**
 - **It also depends on a software tools and usual data size in a given sector.**
 - **„... as technology advances over time, size of datasets that qualify as big data will also increase.“**

Are data more
valuable than oil?



Three approaches (PWC, 2019)

- Market:
 - Active markets for data are rare, mostly illegal.
 - Shutterstock, Flickr.
- Cost:
 - Straight-forward, how much does the data currently cost (e.g. CPC).
 - Fails to capture future revenues a holder can get from the data.
- Income:
 - Measure of cash flows the data are expected to generate.

Are data more
valuable than oil?
It depends on the
data.



How to do research with Big Data?

- The distinction from „normal“ research is in the data collection.
- → How to collect „Big Data“?
 - Google – Trends, Keyword Planner, 3rd parties – SEMRush, Keywordtool
 - Social media – Twitter API, scrapers (Octoparse)
 - Wikileaks
 - Pastebin
 - Cyber security – Shodan (academic licence)
 - Open science repositories
 - <https://openscience.muni.cz/>
- European legislation on open data and the re-use of public sector information
 - <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information>

- How is Big Data (e.g. searches from Google) different from „conventional“ survey/interview/experiment etc. data?



There are pros as well as cons (Davidowitz, 2015)

- Overcome respondent bias (social desirability).
- Efficiency. Wider and deeper insight.
- Representativeness?
- Population of searchers? How big is it?
- Misformulated seed words.
- We need to interpret results with explicit limits and deliberation in the relation with quantitative and qualitative methodologies.



Legitimate ways of use

- Army and law enforcement recruitment (see Jahedi, Wenger and Yeung, 2016).
- Studies on public perception (Kostakos, 2018).
- And others...



- Cambridge Analytica (see Isaak and Hanna, 2018) – Facebook data.
- Bulk surveillance (privacy vs. security debate) – e.g. PRISM programme exposed by Edward Snowden.
- Wikileaks.

References

- Manyika, J. et al. (2011). Executive summary: Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*. Available from: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-Innovation>.
- PWC. (2019). *Putting a value on data*. Available from: <https://www.pwc.co.uk/data-analytics/documents/putting-value-on-data.pdf>.
- Jahedi, S., Wenger, J. W. a Yeung, D. (2016). Searching for Information online: Using Big Data to Identify the Concerns of Potential Army Recruits. *RAND Corp*. ISBN: 978-0-8330-9414-8. Available from: https://www.rand.org/pubs/research_reports/RR1197.html.
- Isaak, J. and Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer* 51(8), pp. 56-59. DOI: 10.1109/MC.2018.3191268. Available from: <https://ieeexplore.ieee.org/abstract/document/8436400>
- Davidowitz, S. S. a Varian, H. (2015). A Hands-on Guide to Google Data. pp. 9-25. (dostupné přes Google Scholar).
- Kostakos, P. (2018). Public Perceptions on Organised Crime, Mafia, and Terrorism: A Big Data Analysis based on Twitter and Google Trends. *International Journal of Cyber Criminology*. 12(1). pp. 282-289. DOI: 10.5281/zenodo.1467919.

Thank you for the
attention. Questions and
your presentations.