Big Data and Security

Moderní technologie a bezpečnost (BSSn4411)

Modern technologies and conflict (CDSn4003)

Jan Kleiner

20.10.2021

MUNI

FSS

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)

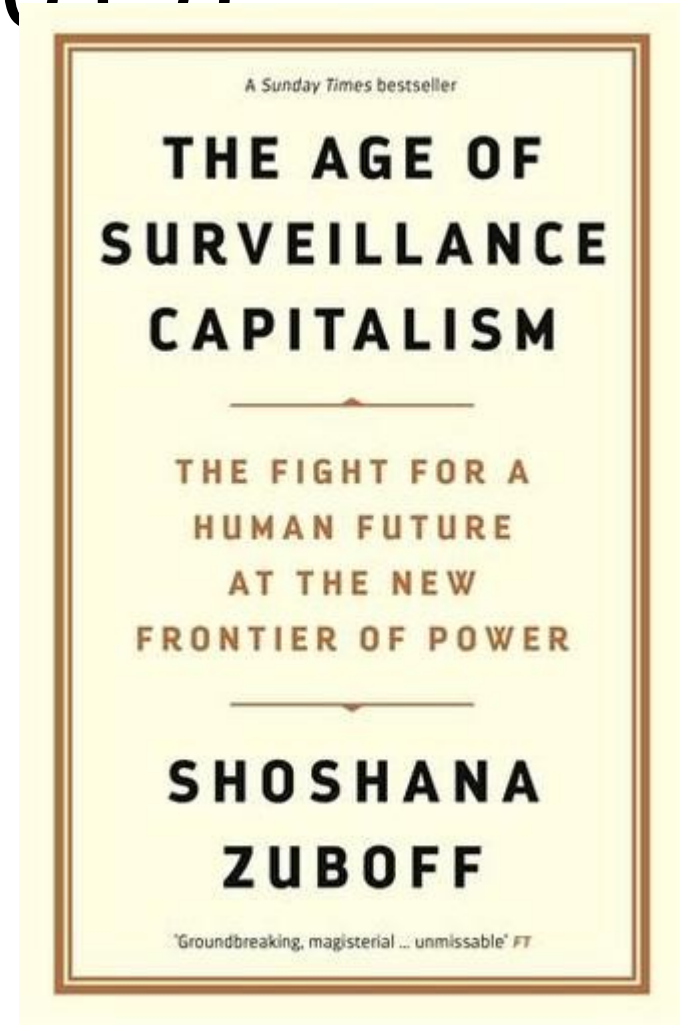
- Market:
  - Active markets for data are rare, mostly illegal.
  - Shutterstock, Flicker.
- Cost:
  - Straight-forward, how much does the data currently cost (e.g. CPC).
  - Fails to capture future revenues a holder can get from the data.
- Income:
- Measure of cash flows the data are expected to generate.
- Around 2017 – Amazon, Google, Facebook – biggest net profits (mainly from advertising).

Surveillance Capitalism (Zuboff, 2019)

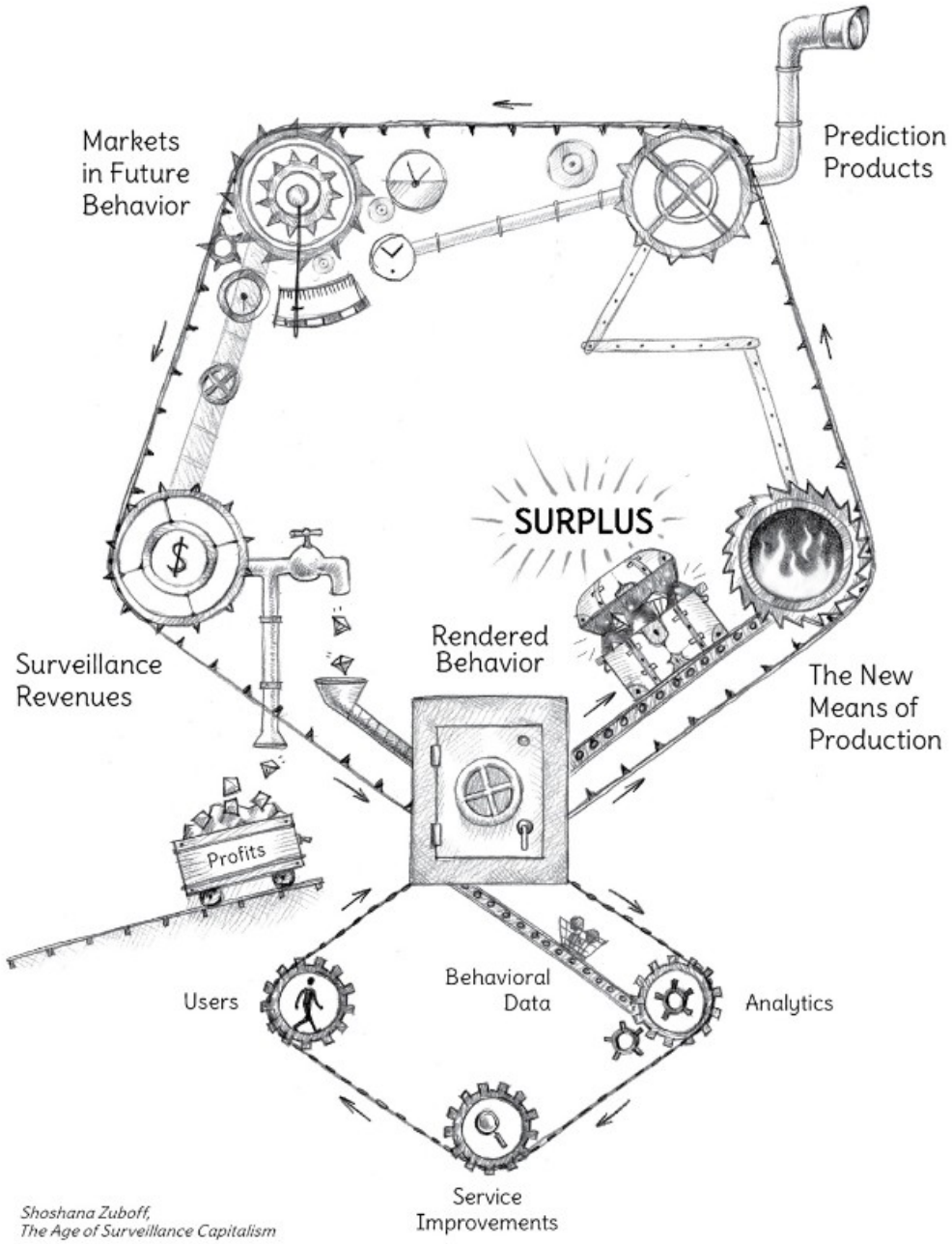
THE DEFINITION

Sur-veil-lance Cap-i-tal-ism, n.

1. A new economic order that claims human experience as free raw material for hidden commercial practices of extraction, prediction, and sales; 2. A parasitic economic logic in which the production of goods and services is subordinated to a new global architecture of behavioral modification; 3. A rogue mutation of capitalism marked by concentrations of wealth, knowledge, and power unprecedented in human history; 4. The foundational framework of a surveillance economy; 5. As significant a threat to human nature in the twenty-first century as industrial capitalism was to the natural world in the nineteenth and twentieth; 6. The origin of a new instrumentarian power that asserts dominance over society and presents startling challenges to market democracy; 7. A movement that aims to impose a new collective order based on total certainty; 8. An expropriation of critical human rights that is best understood as a coup from above: an overthrow of the people's sovereignty.



Behavioural Surpluss (Zuboff, 2019: 97)



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), Facepager
  - Wikileaks
  - Pastebin
  - Cyber security – Shodan (academic licence, shodan trends)
  - 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 intepret results with explicit limits and deliberation in the relation with quantitative and qualitative methodologies.

6 principles of scientific method

1. Empirically testable (through observations, data etc.)
2. Replicability.
3. Objectivity.
4. Transparency.
5. Falsifiability.
6. Logical consistency/coherency.

Challenges (Chen, 2018: 19-23)

- Complexity
  - There is never too little data, only too little processing and analytical power.
  - Social media – complex language issues (e.g., sentiment analysis), enormous scale of data.
  - Data integrity? Not reliable due to lack of accessibility.
  - Transparency? Black-box data algorithms.

- Big Data search
  - Keywords return too many results.
  - The need of post-processing, indexing strategies.
- Lack of theoretical or scientific foundations for Big Data use in research → the need for huge justification (see next slides).
- Other risks (caveats):
  - Fake news, disinformation campaigns/psyops tarnishing.
  - Technology as a catalyst for human behaviour.

- „Big data do not constitute a panacea, and their dark side should never be ignored.“ (Chen, 2018: 22)

How to check for Big Data validity and reliability?

- Measurement (construct) validity
  - Covergent validity
    - Measures of the same trait using different methods show agreement.
  - Discriminant validity
    - Different traits measured by the same method do not agree (any issues here?).
- Multi-trait Multi-method Matrix
- Test-retest reliability (repetition).

	Propaganda perception experiment	from Facebook discussion	Pizza perception experiment	Pizza Facebook
Propaganda perception experiment	=	+++	+/0	0
from Facebook discussion		=	0	+/0
Pizza perception experiment			=	+++
Pizza Facebook				=

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. (via Google Scholar).
- Kostakos, P. (2018). Public Perceptions on Organised Crime, Mafia,and Terrorism: A Big Data Analysis based onTwitter and Google Trends. *International Journal of Cyber Criminology.* 12(1). pp. 282-289. DOI: 10.5281/zenodo.1467919.
- Chen, S. (2018). *Big Data in Computational Social Science and Humanities.* Springer. DOI: 10.1007/978-3-319-95465-3.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight For a Human Future at the New Frontier of Power*. London: Profile Books.

Thank you for the attention. Questions and your presentations.