Big data– Risks and side effects

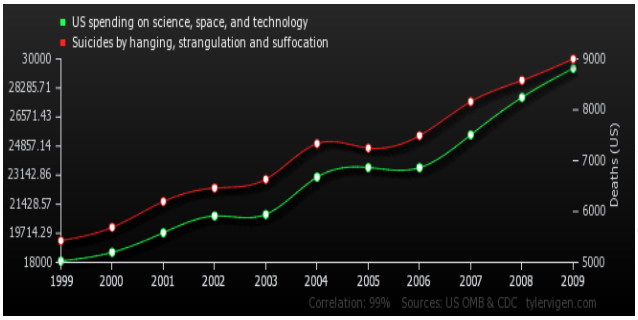
In brief

- Big data is among the most recent technology hypes that promises to win new insights and enhance decision-making in a variety of contexts.
- This kind of technology-aided data analysis is framed as panacea to handle complexity, reduce uncertainty and predict future events.
- Big data bears potential for model-based learning. Tapping this potential requires a deeper understanding of its functioning and its perils.
- Big data can increase complexity and automation which may trigger unintended societal events. Reducing its risks requires transparency, accountability and verifiability of big data analysis.

The potential of big data

Potential applications range from business process optimization, demand-driven energy supply, trend forecasting, medical research, health management, up to predictive policing, etc. Thus, big data surely can support the well-being of society. The general claim that massive amounts of data offer valuable information follows a “the bigger the better” logic suggesting to consider the whole haystack as a gold mine instead of just the needle. The perception that data quality is less important and finding correlation is key for better decision-making is also widespread. Exploring correlations can be beneficial in a number of contexts, particularly for medical research e.g., by showing yet hidden interrelations between symptoms of different diseases, revealing yet unknown patterns that can be very supportive for diagnosis, revealing side effects of drugs, tailored treatment modalities, preventive medicine etc. This can support medical treatment and benefit forecasting and early warning; analysis of anonymized population health data can contribute to explore how diseases (e.g. cancer) develop over time etc. However, this potential is vague and tapping it requires a critical reflection on the claims of big data to come to a clearer understanding of its prospects and limits.

Uncertainty and time limits: Big data can be seen as a noble attempt to reduce uncertainty, often promoted as means to predict future developments. However, its illustrative capacity to point out connections between items is usually based on probabilities and not causalities. The image below shows a correlation of nearly 100% between the US spending on science, space and technology and suicides by hanging, strangulation and suffocation:



Big data trap: spurious correlation

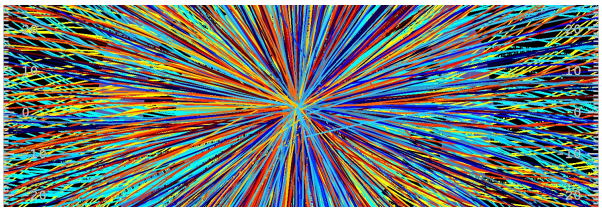
This correlation is obviously complete nonsense and highlights the misleading power of spurious correlations. Correct understanding of big data results is not always as simple as here. Mixing up correlation and probability with causation can boost uncertainty. Big data enthusiasts claim that inaccurate analysis results could be compensated by “big enough data” suggesting that facts come with quantity. Believing this claim not merely risks some inaccuracy but also taking the wrong decisions.

It is crucial to consider that, no matter how “big” it is, data is always a temporal construct emerging in the course of time. Data existing in the present can only provide solid information about the past until the present. In other words: the future remains unpredictable. This simple fact seems to be neglected by big data enthusiasm. Of course, information about the existence of data, correlations, etc. can be very supportive to understand what has happened and what is going on. However, making use of this kind of information requires knowledge about its significance and its limits.

A crucial issue is that complexity increases with large data sets. Together with automated analysis it can become seriously complicated to interpret the information provided by big data, as the following examples highlight.

The big complexity trap

Google flu trends, a seemingly "big" success already demonstrates eventual complications as it overestimated the prevalence of flu by more than 50%. A more serious case is e.g. the faulty calculation of the (Google-related) US big data company 23andme that checks genetic data against health risks. A customer was informed about two mutations in his DNA typical for a genetic disease called limb-girdle muscular dystrophy which is mostly lethal. The customer found a significant error in the analysis and confronted 23andme. They briefly confirmed and apologized via e-mail. In another case, big data was used to support medical diagnosis, where unusual lupus symptoms and an association with a certain propensity for blood clots were found. Based on that, anticoagulant medication was given and the patient did not develop a blood clot. However, this does not prove that there even was a factual risk. It is simply impossible to find out whether the big data diagnosis was a success and the medication correct or the analysis bogus. Such cases highlight that big data can boost complexity and proneness to errors.



Source: wikimedia CC-BY-SA-3.0

Big data visualisations suggest profound findings

Correctly interpreting big data results and its validity can often be far from being trivial. Besides the high mental stress caused by e.g. wrong diagnosis the question is pressing what if such errors remain hidden? Verification or falsification of big data can be complicated. In particular then, if a predicted event is taken for granted and preventive actions are taken. For instance: can a pre-crime be prevented? Or can a merely predicted disease be effectively treated? Thus, big data entails high risks of self-fulfilling prophecies. Trends towards automated decision-making may even lead to situations where the social and economic costs for correcting errors become higher than those for simply accepting faulty big data results. Hence, big data to some extent even seduces to hazard errors. Related are risks of new technology dependencies deeply affecting individual and societal autonomy. Autonomy gains are possible by e.g. early health risk detection, new governance options etc. However, the human factor becoming a subject of statistics strains power asymmetries, privacy and autonomy. Automated predictive analytics might challenge to realize the red line between appropriate intervention and excessive pre-emption. Additionally, individuals could become discriminated because of their DNA. People with risks for genetic diseases then have lower chances for societal development and well-being.

What to do?

To reduce the risks, its likely reasonable reconsidering the thin line between overestimated expectations and underrepresented momentums of uncertainty that correlate with big data. Big new challenges ahead include:

- *Foster data quality and interpretation:* Strengthen analytical skills to handle predictive analytics with care, as humans need to interpret correctly, uncover failure and use results appropriately.
- *Reduce risks of automated false positives*: focus on the information process and its components, e.g., what information in which quality is used for big data, from which sources, for what purpose was it collected.
- *Technology usability:* human computer interfaces that facilitate the handling and interpretation of complex data sets without reducing too much information. This is crucial to allow for scrutiny of big data as well as
- *Accountability, replicability and verifiability* of big data analytics, particularly regarding predictions to reduce risks of autonomy loss and avoid automated decisions.

Further reading

Strauß, S. (2015): Datafication and the Seductive Power of Uncertainty - A Critical Exploration of Big Data Enthusiasm. Information, Bd. 2015 (6), S. 836-847. http://www.mdpi.com/2078-2489/6/4/836

Contact

Stefan Strauß

E-mail: tamail@oeaw.ac.at

Phone: +43(1)51581

IMPRINT: Owner: Austrian Academy of Sciences; Legal person under public law (BGBl. 569/1921 i.d.F. BGBl. I 130/2003); Dr. Ignaz Seipel-Platz 2, A-1010 Vienna; Editor: Institute of Technology Assessment (ITA); Strohgasse 45/5, A-1030 Vienna; http://www.oeaw.ac.at/ita/en. | Frequency: The ITA-Dossiers are published irregularly and publish the research results of the ITA. The ITA-Dossiers are published in print in small numbers only and are made available open access to the public via the Internet portal "epub.oeaw": http://epub.oeaw.ac.at/ita/ita-dossiers | ISSN: 2306-1960 | This work is licensed under a Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0/ |
ITA-Dossier no.22en, May 2016 | http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier022en.pdf