Innovation and Big Data

H. Herik

Innovation and Big Data

H. Jaap van den Herik

Leiden Centre of Data Science, Leiden University, Leiden, the Netherlands
jaapvandenherik@gmail.com

Abstract. Innovation is an essential issue for companies and educational institutes. We have seen that well-known companies came into trouble since they did not innovate in time. They kept their well-established way of operating and adhered to their old-fashioned business models. The main topic of this contribution is to address the question: to what extent can the availability of Big Data support the innovation attempts of companies and educational institutes?

We start by mentioning five examples of innovation and relate them to data science and big data. We describe basic concepts and developments. We then formulate six possible classes of obstacles and take into account the use of sensitive data. Finally, we arrive at two conclusions and three recommendations.

Keywords: big data, obstacles, sensitive data, safeguards, turn around management, narrative science.

1 Introduction

The current technological development is in a state of flux. New applications are followed by even newer ones. Many scientists see this concatenation of applications as a disruptive chain. For a proper understanding we mention five new developments (or applications of well-known concepts): (1) autonomous cars, (2) blockchains, (3) crowd sourced online dispute resolutions, (4) Airbnb, and (5) Uber Taxi. They can be seen as new paradigms that result from the combination of new and old technologies.

The question to what extent big data does support the innovation can be answered by investigating the string of activities that takes place in the study of data science. Data science is a new discipline in the academic world and in particular in the research world. It originates from computer science (informatics) and statistics. Big Data has mitigated the use of samples as is exploited by statisticians. This implies that for a proper handling of the available data no longer the well-known statistical sample techniques are sufficient, but that a new statistical approach has to be developed, i.e., reasoning in a Bayesian network.

In data science we nowadays distinguish seven phases of activities. They are: (1) collecting data, (2) cleaning data, (3) interpreting data, (4) analyzing data, (5) visualization of data, (6) narrative science, and (7) the emergence of new paradigms. The last phase has been instantiated by the five examples at the beginning of this section.

Below we will discuss the following topics and their obstacles: Real big data (section 2); Definitions (section 3); Small data (section 4); Turn around management (section 5); Criminal behavior (section 6); Conclusions (section 7), and Safeguards and recommendations (section 8).

2 Real Big Data

There are only two research areas where we face real big data, viz. particle physics (e.g., at Cern) and astrophysics (e.g., at Lofar). Two other areas where the storage of data is extremely large are DNA-research and Geo-research (in general). For social sciences we see enormous amounts of data stored in Wikileaks, on Dating sites, and in the Panama papers. Here we meet our first class of obstacles. They consist of transport problems, complexity problems, and reputation damage:

Obstacles (Class 1)

Big Data	**Obstacles**
CERN LOFAR	1. Transport problems (accessibility)
DNA GEO	2. Complexity problems (unrelated data)
Wikileaks Dating sites Panama papers	3. Reputation damage (side-effects / non anticipated)

3 Definitions

Big Data is a kind of container concept. Almost all disciplines have their own definition of Big Data. In this contribution we provide two definitions: a technical one and a business one. The technical definition is by Tom White (2012).

Technical Definition: "Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand databases management tools or traditional data processing applications."

The challenges capture

1. curation,
2. storage,
3. search,
4. sharing,
5. transfer,
6. analysis,
7. visualization,

8. interpretation,
9. real-time (van Eijk, 2013).

The second class of obstacles is defined by the characteristics of the collection of Big Data at hand. They are called the five V's:

Obstacles (class 2)

The five V's are:
1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value

The business definition of Big Data has been formulated by James Phillips (2015).

Business Definition: "Big Data focuses on value creation, viz. to obtain operational benefits from data, and will then exploit the benefits for performance improvements." Two important capabilities are:

(1) Companies learn faster on their business trend,

(2) Companies can rely faster on the new trends than their competitors.

The third class of obstacles is defined by legal requirements and competitive behavior:

Obstacles (class 3)

- The intention to make *infringements on the legal requirements* with reference to the commercial competition.

Examples come from the auto industry, among them Volkswagen.

4 Small Data

Next to big data there is small data. They constitute techniques which enables effective use of Big Data. A telling example is *cookies*. Originally, cookies were meant to bring relevant input to places where they are in a better environment, i.e., to be of better use in their "ecosystem".

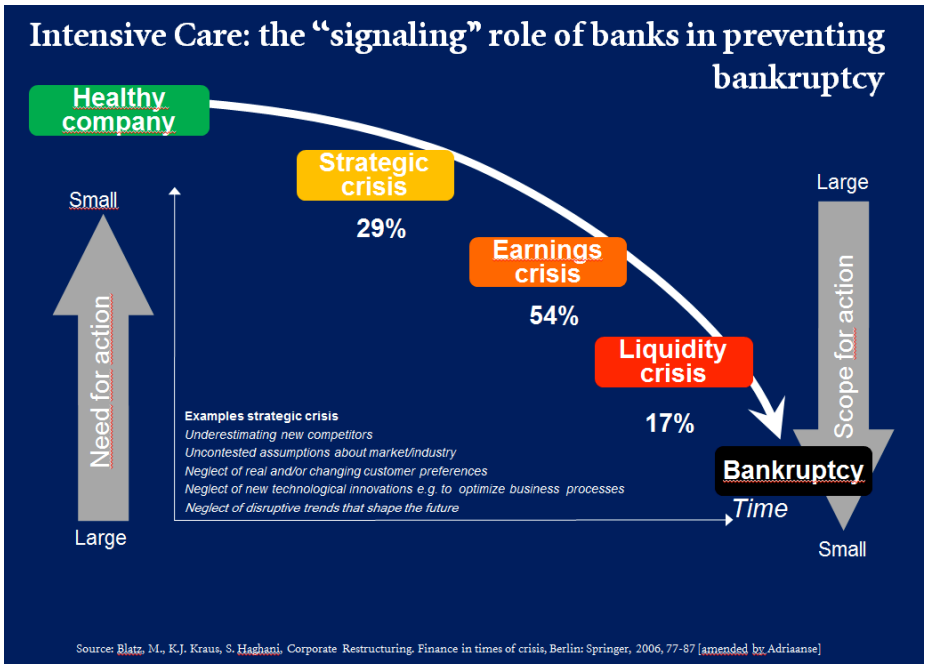
Nowadays cookies are used for: personalization, profiling, advertisements, recording of behavior and for "selling purposes". The fourth class of obstacles is defined by privacy and security issues:

Obstacles (class 4)

The main obstacles for small data are
- Privacy
- Security

5 Turn Around Management

Big data can be used for monitoring the life cycle of companies. We assume that the life cycle traverses through the following six stages: (1) Foundation of the company, (2) Establishing a healthy company, (3) Strategic crisis, (4) Earning crisis, (5) Liquidity crisis, and (6) Bankruptcy (cf. Adriaanse, 2015). We will see the following development



A proper recording of data within the company may serve as monitoring system and even as alert system for unexpected developments. The availability of big data will make the prediction very accurate.

The fifth class of obstacles is defined by privacy and ethical considerations as well as by commercial competition and legal opportunities:

Obstacles (class 5)

The main obstacles for use of big data in Turn around management.

- Privacy
- Ethical considerations
- Commercial competition
- Legal opportunities

6 Criminal behavior

An important case showing criminal behavior is the marathon bombing in Boston. We illustrate this marathon bombing by three pictures. In April 2014 an unexpected bomb exploded at the end of the Boston Marathon (see picture 1). There were sensor images and recorded pictures. The crowd was so big that no criminals involved in the attack could be identified. The police quarantined the city and attempted to find the criminals in their computer databases. They could not identify them. Finally, one brother was killed by police force (Dzjochar Tsjarnajev) in a gunfight and one day later Tamerlan Tsjarnajev was arrested albeit by a coincidence through a blood trace observed by an alert citizen.

If the bomb had been inspected accurately, it could have been established that its origin was Chechnya or a neighbor area, which would have refined the number of candidate criminals considerably. Once the criminals were caught, the recorded images afterwards produced evidence for what has happened and where the guilty persons stayed during the explosion

To investigate such criminal behavior with modern technological means we see that a combination of Big Data, High Performance Computing, and Deep Learning is most effective. These three components form the basis of *narrative science*. It means that an intelligent program constructs the story and points to the criminals. This approach is also known under the name story telling. For lawyers it is called argumentation theory. The sixth class of obstacles is defined by privacy, public safety, technicalities of narrative science, and sensitive data:

Obstacles (class 6)

- Privacy
- Public safety
- Narrative science
- Sensitive data

The investigations of criminal behavior have an important obstacle class 6 to take into account, viz. sensitive data (cf. Van Eijk, 2015):

Sensitive Data

Obviously, some data are really personal, i.e., data revealing
(1) racial or ethnic origin,
(2) political opinions,
(3) religious or philosophical beliefs,
(4) trade-union membership, and
(5) the processing of data concerning health or sex life

The general rule for companies is that the processing of sensitive data is **prohibited** without explicit consent (Directive 95/46/EC, article 8).

The "scientific" development of "disruptive" technology, the Road to Deep Learning, is:

The Road to Deep Learning

Artificial Intelligence	1950-1990
Machine Learning	1990-2000
Adaptivity	2000-2005
Dimension Reduction	2005-2010
Deep Learning	2010-2015
Big Data & HPC	2012-2017
New Statistics	2014-2019

7 Conclusions

From the above line of reasoning on the obstacles that occur when Big Data research is involved we may conclude the following:

Conclusion 1: For a legitimate commercial interest, a legal foundation is a necessary requirement. Such a requirement should constitute a careful balance between

- the legitimate commercial interest,
- the fundamental rights and liberties of the persons who possess the data (or to whom the data can be attributed). Keep here in mind the sensitive data.

Conclusion 2: For all who are working in the area of Technological Innovation and Social Innovation the Extrapolation as shown in below is relevant (cf. Van den Herik and Dimov, 2011).

7 Safeguards and Recommendations

To keep the delicate balance mentioned above, two safeguards are possible.

- To diminish attention and research efforts for Big Data and Deep Learning.
- To increase attention and research efforts for Big Data and Deep Learning.

From these two safeguards our preference goes to the second safeguard, provided that we are allowed to introduce three specific safeguards.

These are our *recommendations*.

- Increase research on AI systems for Big Data and Deep Learning with emphasis on moral constraints.
- Increase research on AI systems for Big Data and Deep Learning with emphasis on the prevention of AI systems to be hacked.
- Establish (a) a committee of Data Authorities and (b) an ethical committee.

Acknowledgements

This insight overview is the result of my cooperation with many colleagues. The following colleagues are singled out for their contribution to this article. I would like to thank Rob van Eijk, Daniel Dimov, Joost Kok, Aske Plaat, and Joke Hellemons. Moreover, I am grateful to Wladyslaw Homenda from Vilnius for the stimulating discussions we have had and for the invitation to communicate on Innovation.

References

1. Adriaanse, J.: Ethical challenges in turnaround management. Lecture at the workshop Leadership Challenges with Big Data. Turning Data into Business, Erasmus University Rotterdam, June 30, 2015.
2. Phillips, J.: Pass summit 2015 – Microsoft foundation session – Business Intelligence (2015)
3. Van den Herik, H.J. and Dimov, D.: Towards Crowdsourced Online Dispute Resolution. *In Revista da Faculdade de Direito Milton Campos* (eds. Carlos Alberto Rohrmann and Rodolpho Barreto Sampaion Junior), Volume 22, pp. 141-162. Belo Horizonte, Del Rey. ISSN 1415-0778 (2011)
4. Van den Herik, H.J. and Dimov, D.: Towards Crowdsourced Online Dispute Resolution. In *Law Across Nations: Governance, Policy & Statutes* (ed. S.M. Kierkegaard, ass. ed. P. Kierkegaard), pp. 244-256. International Association of IT lawyers (IAITL). ISBN 978-87-991 385-9-3 (2011).
5. Van Eijk, R.: Towards a delicate balance in strategic innovations – Update on privacy legislation. Lecture at the workshop Leadership Challenges with Big Data. Turning Data into Business, Erasmus University Rotterdam, June 30, 2015.
6. White, T.: Hadoop: The Definitive Guide, 3rd Edition. Storage and Analysis at Internet Scale. Publisher: O'Reilly Media / Yahoo Press Pages: 688 (2012)