DIACHRON Workshop on Preservation and Evolving Big Data

Panel on Big Data Quality

Bordeaux 15/3/2016

Mokrane Bouzeghoub
CNRS & University of Versailles, FR

Vassilis Christophides
INRIA Paris & University of Crete, GR)

Angela Bonifati (Université Claude Bernard Lyon 1, FR)
Paolo Missier (University of Newcastle, UK)
Norman Paton (University of Manchester, UK)

Panel on Big Data Quality

Everyday we read that Big Data can increase sales, improve performance of business processes, generate new scientific knowledge, make cities and homes smarter, permit safe navigation, allow decision making in a complex world, make people aware of their health, help justice to understand more on criminal behaviour, and even prevent society from violence and radicalism.

Panel on Big Data Quality

- For these purposes, effort is generally put on data analytics, data mining, machine learning, stats analysis …
- Considering data quality either
  - As a preprocessing task not so specific to big data
  - Or based on the assumption that the volume of data compensates possible deficiencies of this data

→ These are the issues to be discussed in the panel!

Questions

- **Q1.** Big Data are usually considered as a guarantee of the results accuracy of data science techniques. What are however the risks for the value extraction chain
  - if the underlying data is wrong, dirty, incomplete, inconsistent, obsolete... ?
  - if the underlying data analysis algorithms are not reliable, biased in their processing or their assumptions ?
- **Q2.** What are the main quality factors of Big Data? Do these quality factors have the same importance in various data spaces : Corporate transactional data, research data, government data, personal data, social media data....?
- **Q3.** Data quality is an old issue in databases, data-warehousing, statistics and decision making systems. To what extent traditional approaches for diagnosis, prevention and curation are challenged by the Volume Variety and Velocity characteristics of Big Data?
- **Q4.** Quality assessment and improvement may be of high cost which can be balanced with the risk to take decisions on the basis of inaccurate data. How difficult is the evaluation of the threshold under which data quality can be ignored ?

Beyond quality …

- Big Data covers different data spaces
  - Corporate transactional data, research data, government data, personal data, social media data
- This data is accessed through web services which may have their own issues
  - Arbitrary filtering of the data provided
  - Introduction of bias and malicious values in the exchanged data
  - Lack of reliable documentation on the provenance and the quality of the data sets
  - Non-ethical conclusions derived from incomplete and imprecise data sets or dishonest sampling
  - Non-ethical purpose for which personal data is used for
  - query answers may be computed wrt business advantage of the service provider instead of only user requirements and preferences. ….

… Data Trust

- What about moving from the data quality concept to data trust concept which can be seen as a global virtue and an ethical attitude driving big data processes and digital economy?
  - Fairness in use
  - Responsability
  - Transparency
  - Ethical behaviour …

→ **Q5: Is the data trust the main quality dimension of Big data ?**