Big Data Round Table

March 22, 2016 (17:00–18:00)

#AgroStat2016

Chairman: Prof. Dr. Diego Kuonen (@DiegoKuonen), CStat PStat CSci,
CEO, Statoo Consulting & University of Geneva, Switzerland

Dr. Mark Wolff, Chief Health Analytics Strategist, SAS Institute, USA

Christophe Sauce, Biometrics Director, Danone, France

‘The four Vs’ provide a definition of big data

- ‘Volume’: ‘data at rest’, *i.e.* the amount of data ($\rightsquigarrow$ ‘data explosion problem’), with respect to the number of observations ($\rightsquigarrow$ ‘size’ of the data), but also with respect to the number of variables ($\rightsquigarrow$ ‘dimensionality’ of the data);
- ‘Variety’: ‘data in many forms’, ‘mixed data’ or ‘broad data’, *i.e.* different types of data (*e.g.* structured, semi-structured and unstructured, *e.g.* log files, text, web or multimedia data such as images, videos, audio), data sources (*e.g.* internal, external, open, public), data resolutions (*e.g.* measurement scales and aggregation levels) and data granularities;
- ‘Velocity’: ‘data in motion’ or ‘fast data’, *i.e.* the speed by which data are generated and need to be handled (*e.g.* streaming data from machines, sensors and social data);
- ‘Veracity’: ‘data in doubt’, *i.e.* the varying levels of noise and processing errors, including the reliability (‘quality over time’), capability and validity of the data.

Key principles for big data analytics' success

- **Do not neglect** the following four principles that ensure successful outcomes:
  - use of sequential approaches to problem solving and improvement, as studies are rarely completed with a single data set but typically require the sequential analysis of several data sets over time;
  - having a strategy for the project and for the conduct of the analysis of data ($\rightsquigarrow$ 'strategic thinking');
  - carefully considering data quality and how data will be analysed ('data pedigree'); and
  - applying sound subject matter knowledge ('domain knowledge' or 'business knowledge'), which should be used to help define the problem, to assess the data pedigree, to guide analysis and to interpret the results.

‘The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.’

John W. Tukey, 1986

'To properly analyze data, we must first understand the process that produced the data. While many ... take the view that data are innocent until proven guilty, ... it is more prudent to take the opposite approach, that data are guilty until proven innocent.'

Roger W. Hoerl, Ronald D. Snee and Richard D. De Veaux, 2014

Source: Hoerl, R. W., Snee, R. D. & De Veaux, R. D. (2014). Applying statistical thinking to 'big data' problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 222–232.

‘Business is not chess; smart machines alone can not win the game for you. The best that they can do for you is to augment the strengths of your people.’

Thomas H. Davenport, August 12, 2015