Rejoinder to the discussion of:

A review of data science in business and industry and a future view by G. Vicario and S. Coleman

We are very pleased and honored to have the valuable contributions of the discussants who have clearly thought deeply about the issues in our paper. Our rejoinder addresses the comments of the discussants in alphabetical order followed by some general remarks.

Ahlemeyer-Stubbe makes an important point about the ethical responsibility of handling data that is often given for free by users. She mentions data slavery in which people give away their personal data and become slaves to the corporations who use it. Data Scientists need to be aware that data slavery may well be coming to an end with people demanding to be involved in analyses which include their data and issues of data ownership becoming increasingly complex.

We might also mention the problem that through Data Science there is an increasing tendency for citizens to reinforce their opinions by exposure to social and other media that is especially aimed at them. This leads to narrow mindedness in people and a lack of innovation in the business world. Data Scientists must not lose sight of the importance of disruptive innovation and ways to encourage innovation are discussed in Ahlemeyer-Stubbe and Coleman (2018) and elsewhere.

Too much emphasis on predictive analytics can inhibit innovation as analysis is essentially based on old data (even if it is only a millisecond old) and can therefore only build on the current situation. It cannot include upcoming ideas or new users with new patterns of behavior. Equipping the Data Scientist with skills in creative, out-of-the-box thinking encourages innovation, for example, introducing random errors into data or sending out non-targeted advertising to a random selection of users from time to time can lead to unexpected consequences which can then help to bring in fresh ideas.

We noted the importance of Data Science to the appreciation of mathematics and statistics in high school and are pleased to learn from the comments by Bailer and Fisher of the International Data Science in Schools Project, www.idssp.org.

Quoting from the website, IDSSP is:

“a cross-disciplinary project involving an international team of computer scientists and statisticians from the leading professional organizations for both disciplines. The purpose of the project is to promote and support the teaching of Introductory Data Science, particularly in the final years of schooling.”

Early introduction to Data Science is vital and will lead to a much better appreciation of data and competence in managing the digital world of the future.

Bailer and Fisher make an important point about the need for ethics in Data Science study so that the Data Scientist can be fully aware of the public aspect of trust in data and models derived from data.

We agree with Bailer and Fisher's comment that we have not specifically discussed the "critical skills associated with Problem elicitation and formulation at the beginning" of a Data Science project, but disagree with their comments that we have not given due recognition to the "equally critical skills required in Communicating the results at the end data" and that "Several passages in Vicario & Coleman's article about the aim or mission or purpose of Data Science appear to fall significantly short of what Data Science is really about, particularly as it provides contributions that will impact business and industry." Our practical experience of Data Science has fully reinforced our belief in the importance of relating findings back to business and industry, and we feel very strongly that Data Science is the opportunity for statisticians to show how they can adapt complex statistical thinking to real life problems.

Bailer and Fisher note that:

"The principles of representative sampling, particularly probability sampling of a specified population, are a critical component of the generalizability of any proposed predictive model."

This statement is fundamental if it is possible to "design the experiment". But in most of the situations encountered in Data Science, data come from the internet, from sensors and other equipment used to collect information. It is mandatory to clean the massive quantities of data before analysis and in this way to try to avoid bias and misunderstanding in the prediction step, which is usually the final aim of the exercise.

Bailer and Fisher also make an important point, saying that:

"As a science, we believe that Data Science relates to accumulating knowledge about data and its uses."

Ferrer comments that "Data Science has emerged to cope with the so-called data tsunami" and on the need for different approaches to deal with petabytes of data. It is true that the use of the descriptive approaches familiar to statisticians is not feasible; but after pruning the "tsunami" with the use of clustering, of principal components, of discriminant analysis techniques (tools familiar to statisticians), the size of the data set is less prohibitive. It is still possible that a scatterplot will appear as a "black blob" and we agree that statisticians have to find "alternative approaches and analysis methods" as suggested in Hoerl et al. (2014). After all the principle that the size of the sample should be not too small in order to have significant results but not too large for economic reasons does not make any sense when there is abundant data at a low cost. Ferrer reminds us of the importance of statistical thinking to avoid the problems of erroneous conclusions from big data analysis. As statisticians we have been promoting this approach for many years, see for example Coleman, 2013.

Ferrer writes: "As commented by Box (1976), Fisher´s work made clear that the statistician´s job did not begin when all the work was over – it began long before it was started". We would like

to add that the contribution of statisticians is only effective if it starts in the planning phase of the investigation. Or else, it happens as declared by Fisher (1935): “When I’m called in after it’s all over, I often feel like a coroner. I can sign the death certificate - but do little more.”

We agree with the importance of asking questions, especially in the Six Sigma DMAIC approach to projects in which the Measure phase starts with asking questions about the resources (including data) that are needed to address the problem described in the Define phase. Questioning is particularly relevant in healthcare (see, for example Coleman, 2011) where problems are multi-faceted requiring input from many sources including subjective feelings, results of objective tests, societal responsibilities and economic considerations.

The change in paradigm mentioned by Ferrer from Question, Data, Analysis (QDA) to Data, Question, Analysis (DQA) is precisely what the Data Scientist addresses when applying her/his trade in practice. This two-way pull and push was noted in Smith et al (2019) in the context of small to medium businesses (SMEs):

“The ideas for analysis arise in two ways. SMEs set up to capitalize on the growth of data, benefit from the ample ideas from customers giving a pull to the data analytics carried out by the SME. In addition, there is a push aspect to the data analytics in which ideas that arise as a result of the analysis can be offered as an additional new service.”

Ferrer discusses team work and we would like to add the important recommendations given by G.E.P. Box, W.G. Hunter and J.S. Hunter (1978):

- Find out as much as you can about the problem;
- Do not forget non-statistical knowledge;
- Define clearly the objectives and have all interested parties agree;
- Learn from each other: the interplay between theory and practice.

In the new teaching opportunities for Data Science, Kenett points out that we need to think about how we teach as well as what we teach. He mentions the meaning reusable learning object MERLO pedagogical methodology as a useful tool and it is to be hoped that the International Data Science in Schools Project is mindful of the need for innovative methods when teaching Data Science in the final years of schooling.

Kenett mentions, in his discussion, the early paper by John Tukey in 1962. We think that Tukey was a pioneer in Exploratory Data Analysis (EDA); he did encourage the analysis of data sets, also with visual methods, in order to extract maximum information from them. But we think that Data Science is something more than EDA. More competencies are required, especially given the complexity of data encountered nowadays. We welcome the list of interesting research areas for Data Science and Industry 4.0 given by Kenett including data integration and data fusion, causality analysis and compositional data analysis (for an example of the latter, see Coleman, 2018).

Kenett makes the important point that:

“Publications should address the gap between academic research and application needs. With this perspective, realistic problems need to be presented as a justification for theoretical

developments. This is different from using an example as demonstrator of theoretical result". He rightly notes the challenge presented by this reversal in the way publications are usually framed – they commonly have theory illustrated by a practical example. Data Science problems and their first guess solutions can be presented and viewed as a valuable pointer to where further research is needed.

Initiating synergistic meetings between Data Scientists and subject experts, as recommended by Kenett, is to be welcomed and relates to the comment by Secchi about Data Scientists needing to be π and wicket-shaped people, discussed next.

Referring to Secchi's discussion, we first consider his comment concerning data complexity and the high dimension of the vectorial space in which data are immersed. We agree that we have not specifically acknowledged that although in the more straightforward framework, the data are vectors, the *data objects* may have a wide range of newer frameworks. Data can be curves, populations and densities of complex objects (common in medical image analysis due to the advent of new diagnostic devices, as underlined by Secchi), elements of more general Euclidean spaces such as Hilbert spaces where functional analysis methods are very successful, or elements of strongly non-Euclidean spaces, such as tree-structured data objects. And for these new frameworks, Object Oriented Spatial Statistics (O2S2) is highly recommended. O2S2 is a system of algorithms and methods for the analysis of high dimensional and complex data with an important spatial dependence. For a comprehensive and thorough review of O2S2 we recommend paper [6] in the Secchi references. The profile of the Data Scientist fits perfectly with O2S2 as it is an intersection of different disciplines including mathematics, statistics, computer science and engineering.

The second issue that Secchi mentions is data fusion and integration. We have to thank him for mentioning a number of scenarios where structured and unstructured data are acquired and integrated with other pieces of information. In many circumstances the responsibility of carrying out the merging is different at each step of the integration. Therefore, new statistical models and algorithms for integrating different sources of data are unavoidable. Quantifying the uncertainty of the predictions, taking into account joint and individual variability, is extremely important. The vision of the Center for Analysis, Decisions and Society of the Human Technopole (CADS - The Human Technopole Foundation) sums these ideas up perfectly:

"CADS will develop original research at the intersection of computer science, mathematics, statistics, artificial intelligence, and socioeconomic sciences, endowing the Human Technopole with advanced data-handling tools and solutions. The Center is founded on the assumption that big data "don't speak" if they are not properly interrogated through original combinations of advanced analytical tools, computing power, technical expertise, and domain-specific knowledge."

Secchi's extensions are valuable as they bring to light issues around space-time data such as in the workings of the brain and the logical progression of the Variety aspect of big data into data from very diverse sources especially prevalent in medicine with sensors, lifestyle data and emotional content all needing to be fused together. Secchi's wider view suggests the dream explored in early data visualization laboratories where the analyst could swim through a sea of

inter-related multidimensional data looking for patterns, Wilkinson (2005) and Wilkinson and Willis (2008).

Secchi makes an important point that knowledge is not always the final goal. He says:

“Stronger and more reliable tools than correlation are offered by the second, and neglected, term in Data Science: science, the systematic and organized knowledge held by the data scientist. This is especially true in the applications of Data Science to business and industry, where the moving force driving the analysis is the need to solve a practical problem, and not necessarily to expand our knowledge of the universe.”

The idea of T-shaped people morphing into π-shaped people is appealing and even more so as they proceed into cricket-wicket shaped people. A good example of a π-shaped person is given in Verma (2019) where a Data Scientist is combining healthcare skills with data analysis. The interviewee, says: “What I do is combine the information that comes from data that human brains can’t analyze, with the expertise, observations and judgement that only humans possess.”

Steinberg and Aronovich raise important points, some of which are fitting subjects for a panel discussion. Concerning the thought:

“….Note how much broader is this task path than what is commonly seen in a statistics classroom, where students often are given a well curated data set, asked to run a well-focused analysis, and to reach conclusions.”,

We feel that academics have to re-think the teaching of Statistics to their students. Skills necessary for expertise in analyzing real data sets and in the application of statistical and decision-making to real operational scenarios should be the prevalent purpose. Therefore, the students need to be asked to “get their hands dirty”, with tackling real life problems. In this they should, of course, be supported by having access to dedicated software for processing massive volumes of data.

Steinberg and Aronovitch (like Bailer and Fisher) are surprised that we note difficulties in adoption of Data Science findings by company CEOs. This is the status in SMEs; in our review we focused more on SMEs as they are a major part of the business world, employing over half of workers in Europe and providing over half of the GDP (Coleman, 2016). In contrast, the discussants cite large tech firms saying:

“We also think that this happy marriage of Data Scientists and decision makers is especially evident in high tech firms in the IT sector, many of which have thriving Data Science teams”.

We would agree that Data Science is better understood and respected in large companies and also in Government administration. It is recumbent upon these institutions to help publicize the successes (and failures) of Data Science. Case studies, such as those presented at conferences and other meetings, articles in trade magazines and user-oriented books all encourage SMEs to make the necessary investment in time and money to take part in the inexorable rise of the digital world.

One more discordant point of view is dating the term Data Science; Steinberg and Aronovitch date the real uptake of Data Science as a phenomenon from 2013 according to Google trends. The early users that we mentioned were from 1960 and then 1974. It is doubtless true that these early users would not have "envisioned the field we see today", but it is undoubtedly in those years that the term Data Science came into view.

Steinberg and Aronovitch correctly suppose that we assume teaching for Data Science includes a basic block of statistical methods.

In our paper we stressed the importance of knowledge generation, however, Steinberg and Aronovitch make an important point that many applications, especially in business and industry, emphasize accurate and generalizable predictions or effective decisions and these will not necessarily advance knowledge. They note that in automobile breaking: "accurate predictions will be the primary need, even if there is limited understanding of why the data led to that particular prediction". But it is clear that knowing which sensors had the most important influence on the performance of the breaking system and how they were related to other sensors is highly valuable and will determine the quality, reliability and maintenance needs of the different sensors.

Steinberg and Aronovitch make the poignant observation that problems are prioritized in Business and Industry by their usefulness to the company rather than by their intrinsic interest. In our paper we cast doubts over the use of black boxes, however, Steinberg and Aronovitch approve of black boxes. We acknowledge the usefulness of black boxes but note that they can still benefit from investigation, as shown in Capaci et al. 2019.

Torelli queries the impression that we may have given that Data Science has mainly arisen in response to the appearance of Big Data. He prefers the explanation that recent developments in computer science, statistics and artificial intelligence opened the way for new collaboration between scientific communities. We agree that this is a reasonable point of view but still feel that the stimulus for Data Science was mostly due to the digitization that lead to Big Data. The abundance of data in industry and research centers during the last decades (from, for example automotive sensors, image recognition and genetics) demands competences that are difficult to frame in separate rigid contexts and this lead to the rise of Data Science.

We made reference in our paper that some people prefer the plural term Data Sciences rather than the singular and Torelli suggests a similar plurality for different types of Data Scientists. The wide range of competences is why there is a variety of curricula in training Data Scientists in universities. In some curricula Statistics education and training is dominant, in others it is competence in Python or other software that is emphasized, or Machine Learning and Deep Learning may predominate the scholar's route. The final goal should be not only to have people educated in specific subjects, but to develop people who are able to work in an interdisciplinary team, and not in a team of "know-it-alls".

Torelli's third and final point reminds us that it is not so straightforward obtaining case studies from business and industry, where trade secrets and competitive edge work against openness.

We regrettably agree that this could “undermine the virtuous process” of everyone sharing and growing together that we predicted at the end of our paper, but we remain optimistic.

It was very rewarding for us to read through the discussants’ comments and there are some consistent themes about which we now make some general remarks. One is the issue of the Data Science curriculum. Steinberg and Aronovitch note the importance of basics, Bailer and Fisher consider our list of topics rather advanced whereas Kenett adds further methods to those we have included.

Ethics are mentioned by Ahlemeyer-Stubbe and Bailer and Fisher; the importance of communication is stressed, in particular by Bailer and Fisher and Kenett; the requirement of synergies between Data Scientists and subject experts is stressed by Kenett, Secchi and Torelli. We add the comment by Professor Steven Marron which also reinforces Ferrer’s views on Data Scientists specializing in different skills and working in interdisciplinary teams:

“I think it’s time for Data Science to consider the concept of team work. Data Science problems with one person working alone are mostly solved, but for solving big challenges it is necessary to have the work of a team of data scientists with different skillsets. Data scientists must now become more interdisciplinary”. (Marron, 2019)

The issue of black boxes also raised comments from the discussants. Black boxes containing algorithms (as distinct from black boxes containing flight recorder details) are clearly important but we persist in our wariness of their general use. In earlier times, citizens were accustomed to the workings of government and law being like black boxes with experts entering information and making decisions. The digital age has briefly opened up these clandestine workings and attempted to make them transparent to those with time, energy and ability to interrogate them. Indeed in the tech world most established practices have embraced full transparency and the results are overwhelmingly positive. Tech companies tend to be fully open about their algorithms, infrastructure, feature maps and deployment procedures and make their code and practices publicly available for anyone to use and contribute to; opening up and being transparent is not just a good will gesture, it's a strategic decision that helps the industry improve. However, with the increasing prevalence of fake news we are in danger of moving back into a world of obfuscation, not knowing what is really going on and how decisions are made, and we want to resist this as much as possible in Data Science.

The role of knowledge in data science also raised some differences in opinion with Bailer and Fisher noting its importance but Secchi and Steinberg and Aronovitch noting that problem solving is often of greater concern. In our review we discussed Data Science mostly in its practical role of improving Business and Industry but we also celebrate its loftier character as a major force in the creation of knowledge.

References

Ahlemeyer-Stubbe, A, Coleman, SY. (2018) Monetising data – how to uplift your business. Wiley.

Box, GEP, Hunter, WG, Hunter, JS. (1978) Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. Page 14. Wiley.

Capaci F, Vanhatalo E, Kulahci M, Bergquist B. (2019) The Revised Tennessee Eastman Process Simulator as Testbed for SPC and DoE Methods. Quality Engineering, 31(2), pp 212-229. DOI: 10.1080/08982112.2018.1461905.

Coleman, SY. (2011) Six sigma in health care, chapter in Statistical methods in healthcare: planning, delivering and monitoring care, Wiley. Editors: Faltin, F, Kenett, R, Ruggeri, F.

Coleman, SY. (2013) Statistical thinking in the quality movement +/-25 years, the TQM Journal, 26(6), pp 597-605. doi: 10.1108/TQM-06-2013-0075.

Coleman, SY. (2016) Data mining Opportunities for Small to Medium Enterprises from Official Statistics. Journal of Official Statistics, 32(4), pp 849-866. https://www.degruyter.com/view/j/jos.2016.32.issue-4/issue-files/jos.2016.32.issue-4.xml

Coleman, SY. (2018) Analysing activities in a classroom – Remembrances of John Aitchison in Hong Kong with applications to a Service Provider, Austrian Journal of Statistics, 47, pp 47-52. https://eprint.ncl.ac.uk/253697.

Fisher, RA (1935) The Design of Experiments. Oliver & Boyd, Edinburgh.

Hoerl RW, Snee RD, De Veaux RD. (2014) Applying Statistical Thinking to 'Big Data' Problems. Wiley Interdisciplinary Reviews: Computational Statistics, 6, pp 222-232.

Marron, S (2019) in an interview reported in "Mathesia Outlook - Data Science Trends in 2019" available from https://mathesia.com/1400-2/

Verma, S. (2019) 'Are you sure you don't have a background in medicine?' Oct 27th 2019

https://brighterworld.mcmaster.ca/articles/manaf-zargoush/?utm_source=researchgate&utm_medium=native&utm_campaign=brighterworld2019-2020.

Wilkinson, L. (2005) The Grammar of Graphics (2nd edition). Springer-Verlag, New York.

Wilkinson, L, Wills, G. (2008) Scagnostics Distributions. Journal of Computational and Graphical Statistics, 17(2), pp 1–19. DOI: 10.1198/106186008X320465