Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 14; doi:10.3390/bdcc2020014

Article

Data Science in Education, Employment, Research: Data Revolution for Sustainable Development

Fionn Murtagh [1,*] and Keith Devlin [2]

[1] Centre of Mathematics and Data Science, School of Computing and Engineering, University of Huddersfield, Huddersfield, UK; fmurtagh@acm.org

[2] H-STAR Institute, Stanford University, Ventura Hall, 220 Panama Street, Stanford, CA 94305-4101; kdevlin@stanford.edu

* Correspondence: fmurtagh@acm.org; Tel.: +44 1484 473827

Abstract: In Data Science we are concerned with the integration of relevant sciences in observed and empirical contexts. This results in the unification of analytical methodologies, and of observed and empirical data contexts. Given the dynamic nature of convergence, described are the origins and many evolutions of the Data Science theme. The following are covered in this article: the rapidly growing post-graduate university course provisioning for Data Science; a preliminary study of employability requirements, and how past eminent work in the social sciences and other areas, certainly mathematics, can be of immediate and direct relevance and benefit for innovative methodology, and for facing and addressing the ethical aspect of Big Data analytics, relating to data aggregation and scale effects. Associated also with Data Science is how direct and indirect outcomes and consequences of Data Science include decision support and policy making, and both qualitative as well as quantitative outcomes. For such reasons, the importance is noted of how Data Science builds collaboratively on other domains, potentially with innovative methodologies and practice. Further sections point towards some of the most major current research issues.

Keywords: Big Data training and learning; company and business requirements; ethics; impact; decision support; data engineering; open data; smart homes; smart cities; IoT

1. Data Science as the Convergence and Bridging of Disciplines

In [23] at issue are: parallels between astronomy and Earth science data, methodology transfer, and metadata and ontologies characterized as being crucial. The convergence or bridging of disciplines must address "non-homogeneous observables, and varied spatial, temporal coverage at different resolutions". Then, given computational support, "it is the complexity more than the data volume that proves to be a bigger challenge". Further benefits of this Data Science convergence are termed here tractability and reproducibility. There is discussion in section 2 of [23] of the complexity relating to resolution and distributions. In [26], this is also characterized in terms data of encoding. Plenty of work now emphasizes the importance of p-adic data encoding (binary or ternary when p = 2 or 3), compared with real-valued encoding (m-adic, especially when m = 10).

The convergence and bridging of disciplines are emphasized in [23], as follows. "Methodology transfer can almost never be unidirectional. Diverse fields grow by learning tricks employed by other disciplines. The important thing is to abstract data – described by meaningful metadata – and the metadata in turn connected by a good ontology." Further description is at issue in regard to Data Science: "We have described here a few techniques from astroinformatics that are finding use in geoinformatics. There would be many from earth science that space science would do well to emulate. Even other disciplines like bioinformatics provide ample opportunities for methodology transfer and collaboration. With growing data volumes, and more importantly the increasing complexity, data science is our only refuge. Collaboration in data science will be beneficial to all sciences."

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

2. Historical Development of Data Science and Some Contemporary Examples of Cross-Disciplinarity

A short historical perspective that follows is with reference to such disciplines as computer and information sciences, mathematics and statistics, physics, and, implicitly, social sciences. In concluding this description, a key point will be how Data Science encompasses and embraces all of the following: cross-disciplinarity, interdisciplinarity and multidisciplinarity.

2.1. Historical Prominence of Data Science in Recent Times

So many of the origins are due to Chikio Hayashi and others. Consider Hayashi [18], "I will present "Data Science" as a new concept.", followed by a very relevant introduction to the science of data, with this: "Data Science consists of three phases: design for data, collection of data and analysis on data". In Ohsumi [31] the abstract has this: "In 1992, the author argued the urgency of the need to grasp the concept "data science". Despite the emergence of concepts such as data mining, this issue has not been addressed."

In the Preface of [15], it is noted how Data Science arises from the convergence of computer science and statistics, that "gives birth to a new science at its core". That Preface concludes with this: "To take data as a starting point provides a complementary vision of theory and practice, and avoids creating an unfortunate gap between two steps, both of which are essential in any scientific process."

In this very comprehensive overview of data science [5], it is stated (section 3.6) how the "first conference to adopt "data science" as a topic" was the International Federation of Classification Societies (IFCS) 1996 conference, in Kobe, Japan. This was fully consistent with our work as participants, then and now (IFCS 2017, in Tokyo, Japan, also had Data Science in its title). In [34], this point is also made about IFCS 1996 as the first conference with Data Science in its title. It is also stated there, [34], that journal, Behaviormetrika, is "the oldest journal addressing the topic of Data Science", when it started in 1974. Data Science is specified as "an interdisciplinary field that includes the use of statistical methods to extract meaningful knowledge from data in various forms: either structured or unstructured".

In [5] there are additional historical perspectives, with the section heading, "The Data Science journey", and this relates largely to work in the 1960s and 1970s. This includes "information discovery" as a continuing key objective in Data Science. This is a key Data Science orientation also in [13]. The latter emphasizes the "semantic dimension of Data Science", through the information discovery lifecyle, and the "discovery lifecycle in text mining". While also emphasizing cooperation, and cross-disciplinarity, there is this:

"We see the data scientist's responsibility

- in the design of an overarching semantic layer addressing data and analysis tools,
- in identifying suitable data sources and data patterns that correspond to the appearance of structured and unstructured data, and
- in the management of the information discovery lifecycle and discovery teams."

An ever-more important issue is the second here, for example cf. section 3 below, arising from the data sources that are employed. As a summary expression, Data Science is, firstly the integration of data sources and analytical and related data processing methodologies, and, secondly and quite fundamentally, arising from the convergence of disciplines. Convergence of disciplines can be so very beneficial in practice. That is, beneficial in regard to addressing and solving problems, and also in regard to the cooperation yielded by cross-disciplinarity. See section 5, below, for some current discussion on how the problems and challenges to be addressed can and should be, quite naturally, arising out of all aspects of Data Science.

The current era of Data Science can be considered as following previous epochs that gave rise to major digital technology advances, with implications in all social domains. Largely the first epoch (in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

the 1980s) brought about laptop and desktop computers, and the second epoch (in the 1990s) gave rise to the Internet and the World Wide Web.

2.2. Practical Association of Disciplines and Sub-Disciplines

In section 2.3, "What is Data Science", in [5], mention is made of Data Science being centred on the following disciplines: statistics, informatics, sociology and management science. Clearly, as in section 5, [5], there is emphasis on "synergy of several research disciplines" and how "interdisciplinary initiatives are necessary to bridge the gaps between the respective disciplines". This is exciting and not least because of how there is convergence of disciplines or subdisciplines. We may consider, for example, how Digital Humanities can incorporate relevant areas of a few disciplines, how computational psychoanalysis can come to the fore (see chapter 8, "Geometry and Topology of Matte Blanco's Bi-Logic in Psychoanalytics" in [26]). With a major focus on psychometrics, Coombs [6] has chapters that proceed from "Basic Concepts" to "On Methods of Collecting Data", and "Preferential Choice Data".

Now, data is so very central to all of our sciences, and to all aspects of our engineering and technology. Just what data is, is a key theme in [26]. That includes data coding, or perhaps also, this should be termed data encoding. After all, data is measurement. It results from this how most important mathematical underpinning is, in Data Science. Implications that follow include the relevance and importance for new, innovative directions to be followed, and from effective problem-solving. The mathematical view of what measurement means is all important. Even in the discipline of physics, in [26] there is the citation of eminent physicist, Paul Dirac, as to how mathematics underpins all of physics, and how the work of eminent psychoanalyst, Ignacio Matte Blanco, has mathematics being integral to psychoanalysis.

From a major study of Big Data and surveying by the American Association for Public Opinion Research, there is the following [20]: "The classic statistical paradigm was one in which researchers formulated a hypothesis, identified a population frame, designed a survey and a sampling technique and then analyzed the results ... The new paradigm means it is now possible to digitally capture, semantically reconcile, aggregate, and correlate data."

Note is made in [1] about wireless connection data forming a basis for public transport management. Such Big Data sources can be associated with, or even integrated with, personal and social behavioural patterns and activities. There is this small heading in [1], "Better living through data?", followed by a very critical statement: "The other thing I need to declare is that I'm no fan of our contemporary belief that life can only get better the more data we have at our disposal." A response to this would be: Data Science, as the science of data, is everything relating to the path and trajectory connecting data, information, knowledge and wisdom.

In [14], it is stated that "The UK's next census will be its last", with administrative, governmental authorities' data replacing the national census. This is acknowledged: "Collecting the data itself is only half the work. A great deal of effort must go into combining it with other sources, in order to answer real questions." That can be understood as undertaking scientific investigation of such data, and other potentially relevant data. The cross-disciplinarity inherent in that also can, and perhaps must, lead to new interdisciplinary linkages. Arising out of the ending of the national census, as such, is [14]: "The way government counts its people is changing, and it could transform policy".

One issue here has been how mathematics underpins so much, across disciplines, and also in the commercial and in most social domains. A good comment to make is this: many universities in the recent past shut down their mathematics departments and no longer provided teaching in mathematics; and now, this is being reversed, with again university courses being provided in mathematics.

3. Open Data, Reproducibility and the Data Curation Challenge

While generally recognized as so important for innovation in both application outcomes and in regard to analytics and methodologies, Open Data plays a key role, for us data scientists.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

(Information and news about Open Data is well provided by this organisation, Open Data Institute, https://theodi.org)

One major aspect of how Big Data analytics are quite central to Data Science is the increasing availability of open data. In [5], this is associated with methodology too, through "the open model rather than a closed one". The following was central to a presentation (in May 2017 in London, UK) by Dr Robert Hanisch, Director, Office of Data and Informatics, NIST – National Institute of Standards and Technology, USA. Dr Hanisch worked for 30 years on the Hubble Space Telescope (HST) project. (The author, Murtagh, of this paper was awarded a medal in 2016: "Outstanding Contributions to Astrostatistics Award, International Astrostatistics Association. Commendation: For his long time contributions to astroinformatics and related areas in the computational sciences; advancing scientific knowledge in classification theory and image analysis; for his contributions to the success of the Hubble Space Telescope; ... and for his long time efforts in dealing with the statistical analysis of 'Big Data'.") Due to open access to observed data, from our cosmos, Dr Hanisch noted that three times the number of people directly engaged in HST work, were working on HST data. It results that there were three times the benefits drawn from HST data.

It was noted by Dr Hanisch how important a role is played by national metrology institutes. Arising from this was, and is, the importance of reproducibility and interoperability of all of analytics comprising Data Science. Underpinning these very important themes in Data Science work is data curation. Data curation is still a major challenge to be addressed. Mentioned in Dr Hanisch's presentation is the contemporary "crisis" of reproducibility. At issue is to support data management from acquisition to publication, or business or medical or governmental or other deployment. The computing expert will recognize this crucial theme of data curation as associated with metadata and evolving ontologies.

For the latter, i.e., the very important and central role of evolving ontology, research publishing, and research funding, are discussed in this broad and general context in [28]. While there remain challenges to be pursued and addressed, it is a good point to make that astronomy and astrophysics offer interesting paradigms for open data, and, in some ways, for data curation.

Noted in this section, is the discussion in [5] of "the open model and open data". It results from this that multidisciplinarity, that we are also expressing as the convergence of disciplines, is to be aided and facilitated by openness of analytics, data management, and all of Data Science methodologies. By openness of methodology, it is intended here to allow domain experts to both link up with, and perhaps even if feasible, to integrate with all that is at issue in other relevant domains. Thus, this is a plea for openness of Data Science as a discipline due to it being a convergence of disciplines.

4. Integration of Data and Analytics: Context of Applications

At issue in this section is an important aspect or byproduct of the integration of data and analytics in Data Science. This theme is to acknowledge, and to seek to address challenges and other issues, in regard to data and the underpinning or contextual reality of the data. Informally expressed, our data represents reality or the context from which the measurements arose, i.e. the data numeric values or qualitative representations.

An outcome of this is to be the quality and standards of our work as data scientists.

Much, and perhaps all, that is at issue in [17] is very important for all that is under discussion here. This reference, [17], describes the problems of data quality, in the Big Data context, relating to administrative data. Hence data curation is very relevant for reproducibility of analytics. There are implications for analytics: "the fact that data are often not of the highest quality has led to the development of relevant statistical methods and tools, such as detection methods based on integrity checks and on statistical properties [...] However, this emphasis has often not been matched within the realm of machine learning, which places more emphasis on the final modelling stage of data analysis. This can be unfortunate: feed data into an algorithm and a number will emerge, whether or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

not it makes sense. However, even within the statistical community, most teaching implicitly assumes perfect data. Challenge 1. Statistics teaching should cover data quality issues."

Our analytics, should not be a "black box", a term that was informally used in regard to neural networks in earlier times. Rather, transparency should always be a key property of analytical methodologies.

The view offered by Anderson [3], and discussed in [24], quoting Peter Norvig, Google's research director, is that "Petabytes allow us to say: 'Correlation is enough'. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." However, this interesting view, inspired by contemporary search engine technology, is provocative. The author there maintains in a provocative way that: "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

It cannot be accepted that correlation supersedes causation, i.e. that analytics can be automated fully, and thereby obfuscate, or make redundant, data science, health and well-being analytics. As described above, in [12], the case is made for comprehensive information governance, encompassing fully the contextualization of all the analytics that are being carried out. In [27] we discuss quite a good deal of the contextualization of analytics of health and well-being data. In the discussion accompanying this seminal work in statistical perspectives on health and well-being, [2], our contribution to the discussion has the following response from the paper's authors: "We agree with Murtagh that 'big data' may offer insights, provided that there are appropriate analytics."

It is so very relevant to note here that data science has inherent and integral involvement in the sourcing and in the origins of data, i.e. selection and measurement. This point is emphasized for applications in [36]. This implies full integrating of the analytics with what data is selected and sourced, and that may well imply what and how measurement is carried out.

An aspect resulting from this section may be the priority to be accorded to induction-based, i.e. inductive, reasoning (cf. [26]). This could be a minor argument for the importance of approaches that follow from data mining, unsupervised classification, latent semantic analysis, and various other themes. Very clearly, however, all of one's studying and teaching, one's work for companies, for Government agencies, health and other authorities, all one's work should and really must be properly focused on the aims and objectives. The latter, of course, may need, partially in any case, to be determined by the expert data scientist.

5. Short Review of Contemporary Data Science in Education and in Employment

In this section, the two themes are relating to the contemporary context. While comprising a short review here, first there is the theme of higher education in Data Science, and second there is the theme of company employment advertisements. Used here is accessible and available data. Of course, an expert Data Scientist is very likely to be involved in many discussions and debates with current and potential students, and with company executives and with many others, besides. It can even be seen that most disciplines have to be integrated into data science. In this regard, for education, see [7].

5.1. Teaching and Learning for Data Science

Briefly considered here are current higher education post-graduate programmes (usually termed MSc courses) in Data Science. This is also possibly relevant for undergraduate programmes, and certainly relevant for undergraduate projects and company placements of students.

In universities in all countries worldwide, in recent years there has been a very great increase in graduate level courses in Data Science, and increasingly also in undergraduate level courses. In Press [33], there is a listing of graduate courses, in some cases but not in all, with the title Data Science. This listing, with links to the host institute, contains: 102 MSc courses listed, 19 online courses, followed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

then by 11 free online courses, and 8 for a fee. Therefore this has 140, for the most part, graduate level courses.

What follows now is, again, the theme of having data and analytics well integrated.

Considering the most essential requirements of a data scientist, in Englmeier and Murtagh [12], we note the very close linkage between data science and Big Data. We emphasize the great need to avoid false positives coming from the data science analytics that is carried out. This arises from treating the data without fully linking and even integrating the analytics with the context, the relevance, and all that is to do with application and problem conceptualization. Noted in that article are the well known errors arising out of the Google Flu Trends, and service usage patterns obtained from Uber. These were outcomes that produced false positives. There must be fully comprehensive information governance, encompassing all levels of information discovery, through conceptualization that can benefit very much if collectively undertaken.

5.2. Employment Requirements in Data Science

So many employment possibilities are now on offer for a Data Science role. In the very comprehensive review of Data Science, [5] has section 6 entitled "Data economy: data industrialization and services". An increasingly popular web service entitled "DataScientistJobs" is available at `https://datascientistjobs.co.uk`.

In this section, there is a small data analysis carried out of stated requirements for Data Science posts. While one must give the fullest perspective to the companies that one works with, and the university Data Science courses that one teaches, what follows is both a consideration and a selection of data, and preliminary results. This preliminary study of requirements for Data Science roles, some of them in senior management roles, represents what we will further pursue over time, both for the benefit of our Data Science students, and to be fully prepared for our work, and association, with companies, nationally and globally.

Online discussion of Data Science and of Big Data have become commonplace, and there are often surveys carried out. Examples include [30] on Big Data, where senior corporate executives were surveyed, and the dominant sector was financial services. This survey concerned internal investment and organisational matters, and business practices and plans. In [19], more than 620 data professionals were surveyed, in regard to skills required and at issue in Data Science. There is an interesting summary, and presentation, of results obtained in a factor space.

We considered description of new posts as Data Scientist, from 2015 to 2017, all from the distribution list, StatsJobs (sometimes indirectly, through links), in England. In most cases, indicated were languages or software environments that are at issue. In a few cases, the job advertisements do not explicitly list these details. Retained for use here were 73 such job descriptions. The very frequent (more than 4 advertisements) software languages and software environments were as follows, required by the 72 employers here: R (50), Python (44), SQL (30), SAS (28), Hadoop (25), Matlab (17), SPSS (17), Java (16), Hive (14), Excel (9), MapReduce (9), NoSQL (8), Spark (7), C++ (6), Pig (6), Tableau (6), HBase (5), C# (4), Mahout (4), QlikView (4), Scala (4).

To have sufficient comparability of software languages or environments, the above 21 of these were selected that were required by at least four employers from the set of advertisements here. Since some employers had none in this set of software languages or environments, and indeed about three of the set of 72 had no detailing at all about what was required, consequently the set of employers was reduced to 60. Thus we have a cross-tabulation of 60 employers, wanting to employ a Data Scientist, with a few requirements or desires for expertise in the set of software languages and software environments that are listed above. The manner of expression was most often: one or another or another again.

Correspondence analysis takes the employer set, and the software set, in the dual multidimensional spaces, both endowed with the chi squared metric, and maps both clouds into a factor space endowed with the Euclidean metric. Hierarchical clustering was carried out from the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

full dimensionality (therefore with no loss of, or decrease in, information content) factor space. We are seeking just to see what associations of software are most likely to be the case, from these Data Scientist job advertisements.

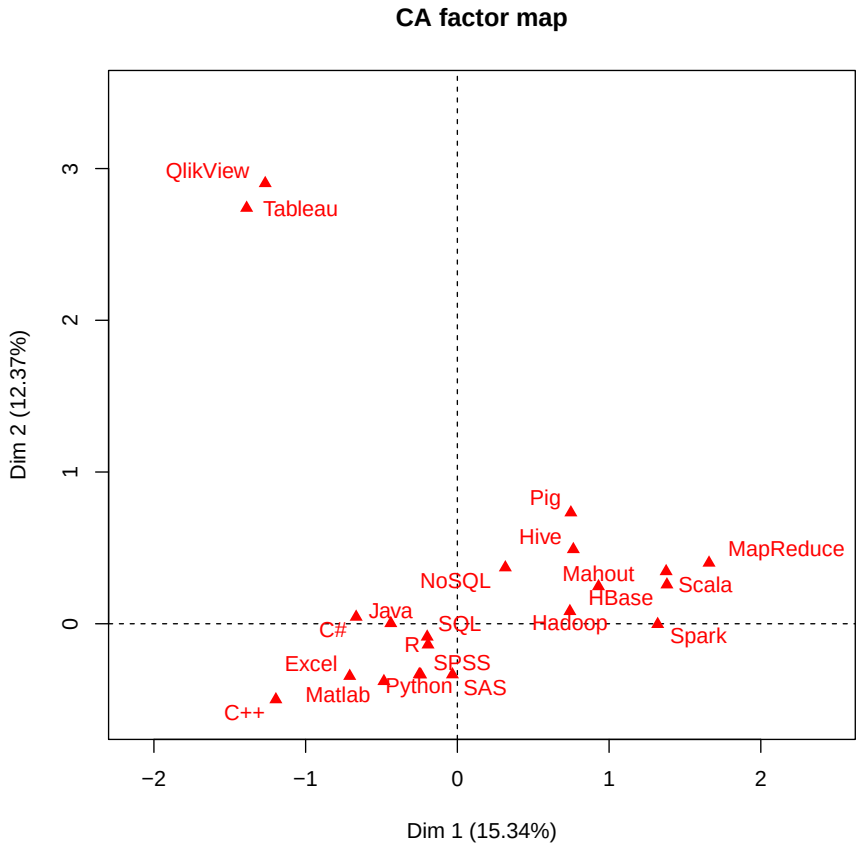

Figure 1. From 60 Data Scientist job advertisements, non-empty from the initial set of 72 employers, with use of 21 software languages or environments. The latter were required by at least four employers. Displayed is the principal factor plane.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

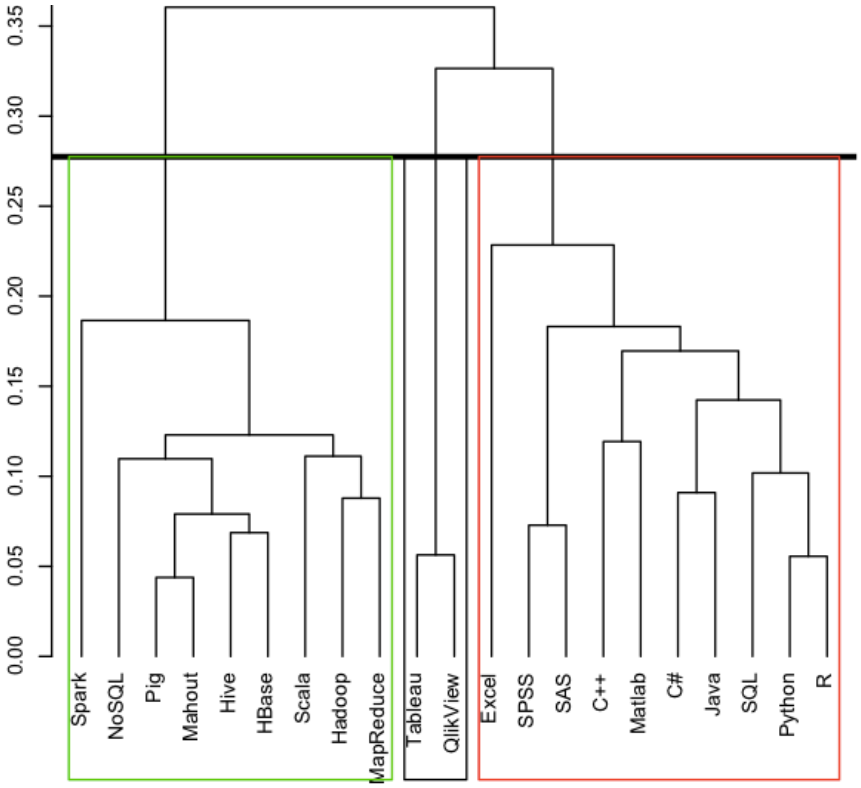

Figure 2. Hierarchical clustering, and derived 3-class partition, of the 21 software languages or environments at issue here, based on the full dimensionality, Euclidean-metric endowed, factor space. See section 5.2. The agglomerative criterion is the Ward minimum variance method.

Figures 1 and 2 display the clustering of the software languages or environments. It is our intention to take such a mapping much further, with supplementary elements, also termed contextual elements, to locate them in the factor space, and these would include: country or location of the job, industrial sector or government agency, or global corporate firm, for the job. The objective is to determine sectoral or regional preferences in the skills and abilities of the Data Scientists employed.

6. Data Science Methodology to Address: Selection Bias, Scale and Aggregation Effects, and Qualitative Evaluation of Decision-Making Impact

Following the pointing to certain challenges in particular in Big Data analytics, involving selection bias and replacement of individual attributes with aggregrated attributes (hence, the collective attributes of groups to which the individual can belong), the main aim here is as follows. It is to point to innovative new methodological perspectives that can both address such issues and challenges, but also benefit from the context, for example of using Big Data sources. In [29] a case study involving work for a major company is illustrated with an example of how aggregated data can be used, if required, for individual-related analysis.

Ethical as well as methodological issues arise in scale effects, representation and expression, and particular context effect. Here we both summarize the ethical implications, and the potential for qualitatively and quantitatively evaluating impact of decision-making and of policy-making.

The quite regular lacking of coordination, alignment and integration of methodology including modelling, with data sourcing, is pointed to in Hand [16]: “ignorance of selection mechanisms has led to mistakes”, “This applies in human interactions – where it has been suggested that the notion that ‘data=all’ can replace the need for careful theorising and statistical modelling – but also in the hard sciences and medicine.”

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

In Keiding and Louis [21], it is well pointed out how one case study discussed "shows the value of using 'big data' to conduct research on surveys (as distinct from survey research)". Limitations though are clear: "Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external . . . pool, in part because of self-selection, ...".

Important points towards addressing these contemporary issues include the following. "When informing policy, inference to identified reference populations is key". This is part of the bridge which is needed, between data analytics technology and deployment of outcomes.

"In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data." While "Representativity should be avoided", here is an essential way to address in a fundamental way, what we need to address: "Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws". In our discussion of the important issues here, in [21], it is noted how, related to eminent social scientist, Pierre Bourdieu's, work with homology between fields of study offer clear perspectives on how beneficial innovative practice can be pursued.

This incorporates our need to "rehabilitate the individual" in our analytics, and not simply replace the individual by the mean of some group, Many case studies of the latter are provided in this book by an eminent mathematical data scientist: O'Neill [32]. From [22]: "Rehabilitation of individuals. The context model is always formulated at the individual level, being opposed therefore to modelling at an aggregate level for which the individuals are only an 'error term' of the model."

Calibrating surveys and other data sources, through use of Big Data, has been at issue in addressing challenges and obstacles described in Keiding and Louis [21]. In regard to decision-making and policy-making, the analysis of discourse in a data-driven way can provide relevant or necessary contextualization. Without having such an approach, there is the following limited capability on the part of those in authority [25]: "top-down communication campaigns both predominate and are advised by those involved in social marketing ... However, this rarely manifests itself through measurable behaviour change ..."

Instead, mediated by the latent semantic mapping of the discourse, we develop [25] semantic distance measures between deliberative actions and the aggregate social effect. We let the data speak in regard to influence, impact and reach. Impact is defined in terms of semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested. It can be visualized. It can be further visualized and evaluated.

For research and for all engagement in Data Science, it is so very motivational to both address, and have significant achievements, in regard to innovative methodology.

7. Benefits of Very High Profiling of Data Science

There are many blog postings, currently, with the theme of "Big Data is dead". (A Google query of the phrase, dated 2017-12-29, lists 153,000 results.) At issue is just this: complete priority is to be given to the problems to be solved and the challenges to be addressed. In the extensive and outstanding detailing of many aspects of Data Science in [5], there is the acknowledgement that there is much that is still currently "tremendous hype and buzz", and "engendering enormous hype and even bewilderment". There is this perspective too, which can be a viewpoint if the sole aim were for Data Science to automate data analytics in all domains of application: counterposed to advanced analytics, "dummy analytics is becoming the default setting of management and operational systems" [5].

Fully in line with context of those perspectives, a major theme of this article is that the convergence of disciplines, in the Data Science framework, builds on cooperative and collaborative expertise, and thus does not seek to replace or supplant such expertise. Thus a major conclusion is not to replace current disciplines (mathematics, statistics, computing, engineering, physics and chemistry, arts and humanities, social and psychological sciences, and so on) but, where relevant and where appropriate,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

and also where motivated and where justified, to re-orientate and to bridge primary as well as foundational levels of disciplines.

In a somewhat humorous fashion, in the sense of revolution versus evolution, let the following be noted. At the 61st World Statistics Congress, in July 2017, in Marrakech, Morocco, there was a session organized jointly by the High Commission for Planning (HCP), Morocco, and the Ministry of Development Planning and Statistics (MDPS) of Qatar. This session was entitled "The Data Revolution for the Sustainable Development Goals". One comment raised in the Question and Answer session was a request for evolution to be at issue rather than revolution. It is also interesting to note how there is an important Advisory Group in the United Nations, called the Data Revolution Group. See: http://www.undatarevolution.org . This has the theme of: "Mobilising the data revolution for sustainable development". For Data Science, it is clear that there is great inspiration here. Some other organisational initiatives will now be mentioned. This is to complement a great deal that is being done, already, by major organisations in statistics, in classification and data research, in engineering, and explicitly in Data Science.

In European research funding, i.e. Horizon 2020, an important supported project is entitled the European Data Science Academy (http://edsa-project.eu). EDSA dates from 2005. There could well be an important role for such an organisation in the future, in regard to sponsoring fellowship levels of organisational memberships, and it would be interesting to promote chartered membership. In the European Commission context, dating from July 2014, there is this: "Best practice guidelines for public authorities and open data" under the theme of "Commission urges governments to embrace potential of Big Data" (`http://europa.eu/rapid/press-release_IP-14-769_en.htm`).

At UK national level, an important initiative, directly or indirectly related to much that was under discussion in this article (in section 3, in particular) is open data. The Open Data Institute (cf. https://theodi.org) in the UK was founded in 2012 by Sir Tim Berners-Lee and Sir Nigel Shadbolt. In welcoming membership applications, there is this: "Membership: Join the data revolution". There is this prominent statement too: "Data is changing our world".

In a practical sense, focused on data to begin with, and no doubt whatsoever, relevant for data curation now and in the future, cf. section 3, there is the Research Data Alliance, RDA (https://www.rd-alliance.org). RDA is supported from the EU, from NSF and NIST in the US, by JISC and other agencies in the UK, and from Australia and Japan.

8. Important New Research Challenges from Data

Data Science, integrating potentially all application domains, with mathematical foundations for methodology as befits observational science, and integrated observational and experimental science, fully relates data to all that is accomplished and achieved from the data sources. This results in the great importance of contemporary increasing orientation towards, and requirement for, open data. The following is a good understanding of this development in Data Science, and of the potential here for application transfers, in parallel with methodology transfers, [23].

The Open Universe initiative (http://www.openuniverse.asi.it) was established by the United Nations, [35]. This work involves: "Today acknowledging that open data access drives innovation and productivity is a well-established principle in every scientific discipline. However, there is still a considerable degree of unevenness in the services currently offered by providers of data...". Among six objectives there is: "Advancing calibration quality and statistical integrity", with outcomes for education, globally, and private sector involvement. Here and, through transference to each and all domains for Data Science, what is required for open data and, in this motivational and inspirational work, open data and all associated open information must be: Findable, Accessible, Interoperable, and Reusable, the FAIR principles, and Reproducible [37].

Supporting the FAIR principles is ESASKY (European Space Agency, Sky), accessible from http://sci.esa.int/home, and described thus: "ESASky, a discovery portal that provides full access to the entire sky. This open-science application allows computer, tablet and mobile users to visualise

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

cosmic objects near and far across the electromagnetic spectrum." The interesting new research challenges in Data Science can be stated to be foremostly related to the transfer to many domains of FAIR-based open science, discovery portals.

Quite an important application domain here will be emerging smart technologies, that encompass smart homes, smart cities, smart environments in general, and Internet of Things technologies. An important Situation Theory methodology is in [8], in an information space that is mathematically based, furnishing a comprehensive representational system. Associated with this are the social, legal and economic aspects of emerging smart technologies in real-life applications.

9. Information Space Theory for Big Data Analytics in Internet of Things and Smart Environments

Context is so very important in Big Data analytics and in many domains, [27].

Situation Theory is to provide humans (generally, trained domain experts) with powerful, flexible representations that enable them to perform better, both as analysts and decision makers. Systems such as the one outlined in Figure 3 for the US Army have a software back-end, possibly including AI, but they are in no way "calculators" or expert systems for making decisions. What was done was to harness the power of mathematics primarily as a representational system, compared to its computational capacity. While the back-end software can manipulate the network – each completion diagram is a structurally identical piece of code – perhaps permitting the eventual application of familiar-looking network-optimization algorithms, many of those completion diagrams represent inherently human thoughts, intentions, and actions, and for the foreseeable future the human mind remains the best tool to handle them. This work for the United States Army was to use Situation Theory to develop a first-iteration specification for a workstation to be used by a field commander, in both mission planning and real-time control. This work includes the taking account of the many different ontologies in a modern battlefield. The role of ontologies is very central in qualitative analysis of research, cf. [28].

9.1. Context, Situation Theory, Completion Diagrams

In the early 1980s, a group of researchers at or connected to Stanford University started to develop an analogous mathematically-based representation of communicating humans, looking deeper than the mere fact of communication (captured by the network model used by the telecommunication engineers) to take account of what was being communicated. (Part of the challenge was to decide how far it is possible to go into categorizing that "what" in order to achieve a representation that is useful in analyzing communication and designing communication-based activities such as work.) That approach is generally referred to as Situation Theory. Devlin was one of those early pioneers, and wrote a theoretical book on the subject, Logic and Information [8]. Subsequently, [9] and [10] apply the techniques developed by the Stanford group to solve an actual workplace problem involving communication in the workplace.

The representation [10] used was (of necessity) similar to that used by telecommunication engineers, Google, the postal system, UPS, FedEx, in that the domain is represented by a network. But whereas those earlier examples had networks of point nodes, the nodes in network were more complicated objects, which were termed "completion diagrams." See to the right hand side of Figure 3, where "situation s1" results in "type T1", and "situation s2" results in "type T2", so that transition from "situation 1" to "situation s2" has the related association between "type T1" and "type T2". The exact nature of the entities in such a completion diagram: they can be considered as capturing the key elements of a basic human act, here military and managerial, including a communicative act. Much of Logic and Information is devoted to the development and explication of such a completion diagram. It has its origins in [4].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

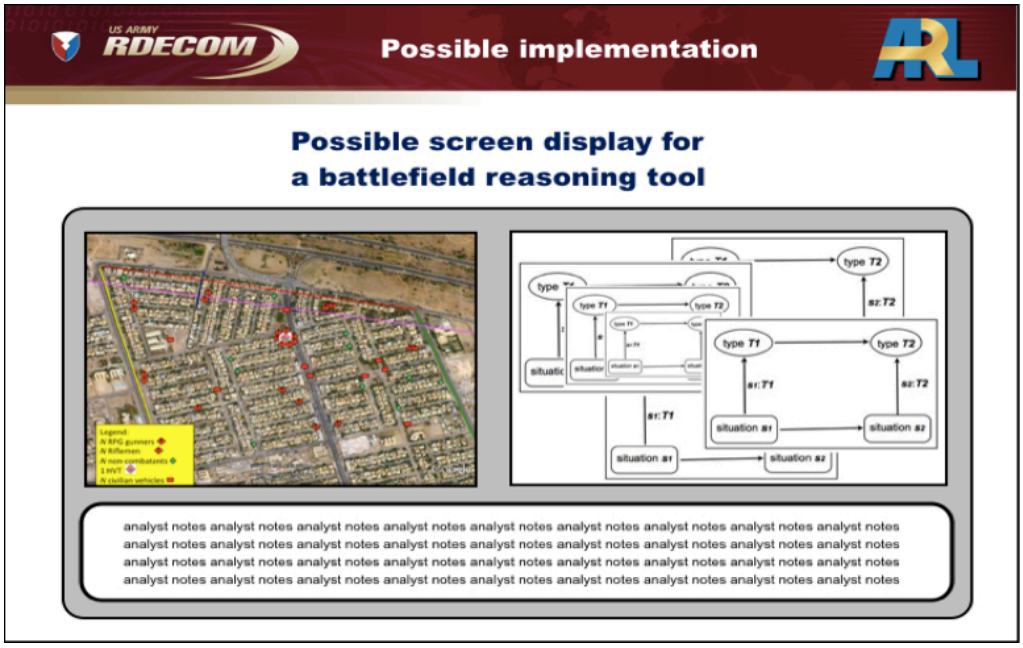

Figure 3. The completion-diagram network is complex. In the screenshot, the aerial view (taken from a previous mission used in training) is of an urban battlefield. The back-end system links the elements in each completion diagram to a corresponding feature in the aerial view, permitting the user to work fluidly with the two representations, having the benefit of two very different views, one spatial, the other human-structural, so the user can explore the domain from (literally) different perspectives.

Information is a vehicle for the use of a Big Data approach to underpin the study of interaction and communication in smart environments (cities, workplaces, homes). Information Space Theory is to provide the focus for building an inter-disciplinary community concerned with social and technological issues associated with recent technological advances. Relevant emerging research and innovation disciplines include Internet of Things, Internet of Everything, Big Data Analytics among others, that contribute to the design, development and effective implementation of Smart Environments in real-life.

The Information Space Theory takes into account the following. (i) People who inhabit smart environments and spontaneously generate data and information in the course of their day-to-day activities. (ii) Place which can be public (smart cities), privileged (workplaces) or private (homes) with varying degrees of privacy and security constraints that shape information sharing. (iii) Patterns of interaction between people and technology that is an integral part of smart environments and influences human-human, human-device and device-device interaction.

A summary follows of Inter-disciplinary Information Space Theory and its application in Smart Environments: (i) Introduction to studies of Information, Data and Interaction. (ii) Big Data Analytics as a tool for the development of Information Space Theory. (iii) Information Space Theory and its impact on the design of Smart Environments. (iv) Information space and human communication research, involving an account of the evolution of smart interaction systems. (v) Further refinement of Information Space Theory informed by cross-disciplinary perspectives and requirements of application in smart systems and emerging technologies, including contribution to the application of Big Data Analytics in real-life smart environments. (vi) Emphasis can be on introducing the concept of Information Space as distinct feature of human context that makes it possible for people to achieve

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

coordination and reciprocity of perspectives through smart interaction systems that safeguard their privacy and security.

Such work is to build on the work of an inter-disciplinary group of researchers within mathematics, computer and social sciences who will together address the key research questions – how do emerging smart technologies influence information sharing in interaction between people and technology in smart environments? What are the social, legal and economic impacts of emerging smart technologies in real-life application?

To this end, the concept of Information Space will guide the investigation into interactions that occur within smart environments taking account of human-human, human-device and device-device in a uniform framework. Special attention is given to information sharing – pathways, enablers and gatekeepers – to incorporate security and privacy concerns that urgently need to be addressed in order to optimise the technology potential in real-life applications of smart environments. The working assumption behind this approach is that inter-disciplinary, formal, theoretical understanding of the nature of these interactions is essential for these concerns to be addressed and resolved.

In this context, mathematics plays a crucial role in developing and using a mathematically-based representation framework for the analysis and design of work in the era of the Internet of Things. Both in life and in scientific studies, what we can achieve depends on, and is constrained by, the representational system we use. The greater the complexity of the domain, the more significant is the representation at our disposal–representations are what make it possible for us to understand and reason about the world. For instance, trade, commerce, and financial activity in Europe were revolutionized by the introduction of the Hindu-Arabic, decimal arithmetic system ("modern arithmetic") in the 13th Century, which made it possible for anyone to become proficient in arithmetic after just a few weeks practice. A similar revolution occurred in the 1980s, when the introduction of the modern, windows-icons-mouse interface for personal computers made it possible for ordinary people to use what had until then been a tool for trained experts. Long before those two examples, the introduction of numbers themselves, in the form of a monetary system, transformed human life by providing a simple, quantitative representation system for property ownership and social indebtedness.

The rise of natural science involved a new representation system that assigned numerical values to various features of the environment (features given names such as length, area, volume, mass, temperature, momentum, etc.) and shifted the focus from trying to understand why things occurred to simply measuring how one quantified feature varied with another – an approach that proved to be extremely fruitful for society. The representation systems of the natural sciences have all been based on mathematics to a considerable extent. In the social realm, mathematically-based representation systems are less common, but when they have been developed they have proved extremely powerful. (Money is a particularly dramatic example.) Indeed, one of the most widespread applications of mathematics in today's world is the optimization of various human activities. Computer search depends on optimization in a mathematical space that treats every living human as a node in a simple mathematical structure called a graph. "Modelling" a person as a point node in a mathematical network omits all information about a person save for one factor: the connections of that human to all other humans. But for questions that hinge on that one factor, the representation enables mathematical algorithms to be applied that provide society with one of its most important tools.

Another example is provided by the algorithms that route our telephone calls, our Internet communication, or mail and package delivery systems, and our transportation systems. In those cases, whereas search engine like Google represents the human domain as a two-dimensional network of nodes and edges, the domains of communicating devices such as phones or computers, of letters and packages in shipment, and of travellers are represented as high dimensional "polytopes," generalizations of the familiar polygons of high school geometry to higher dimensions, to which mathematical methods such as the Simplex Method or Karmarkar's can be applied to determine optimal routings. These representations work by ignoring almost everything about the entities in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

domain apart from the one or two features that are germane to the task. The result is that the power of mathematics can be brought to bear to a problem that, on the face of it, is part of the complex web of human activity that defies the methods of science in terms of its complexity and (local) unpredictability.

10. Conclusion

Having indicated a few highly important, and relatively recent, organisational initiatives (cf. section 7), let us again emphasize that Data Science, viewed as the convergence of disciplines, or, in practice, sub-disciplines, should very much incorporate open methodology, open data, and transparency, reproducibility, and interoperability. (Cf. section 3).

This article has sought to form a foundation for further study of the specific content of Data Science education and training (cf. section 5.1), and of business sectoral importance (cf. section 5.2). After all, progress and impact ensure development and evolution over time. As noted above, too, we may, if we wish, refer to the contemporary data revolution.

Both challenges (cf. sections 4, 6) and impactful potential (cf. section 2.2) are prominent, and it is good to see them as predominant, in our rapidly growing (cf. section 2.1) discipline of Data Science.

There are, in sections 8 and 9, the most important directions to both follow and to incorporate in other domains.

References

1. Abbany, Z.: A public transport model built on open data, News article, (27 Nov. 2017).
   Access: http://www.dw.com/en/a-public-transport-model-built-on-open-data/a-41546053
2. Allin, P. and Hand, D.J.: New statistics for old? – measuring the wellbeing of the UK, Journal of the Royal Statistical Society A, 180 (1), 3–43, (2017). Including F. Murtagh comments.
3. Anderson, C.: The end of theory: the data deluge makes the scientific method obsolete, Wired Magazine, (16 July 2008)
   Access: http://www.wired.com/science/discoveries/magazine/16-07/pb theory
4. Barwise, J. and Perry, J.: Situations and Attitudes. Stanford: CSLI Publications, 1999.
5. Cao, L.: Data science: a comprehensive overview, ACM Computing Surveys, 50 (3), 43:1–43:42 (2017).
6. Coombs, C.H.: A Theory of Data, Wiley, Hoboken, NJ, (1964)
7. Kei Daniel, B.: Reimaging research methodology as Data Science, Big Data and Cognitive Computing, 2(4), 1–13, (2018)
8. Devlin, K.: Logic and Information, Cambridge University Press (1991).
9. Devlin, K. and Rosenberg, D.: Language at Work: Analyzing Communication Breakdown in the Workplace to Inform Systems Design, Stanford: CSLI Publications, 1996.
10. Devlin, K. and Rosenberg, D. "Information in the study of human interaction", in Johan van Benthem et al. (eds), Handbook of the Philosophy of Information, North Holland (2008), pp. 685–710.
    Access: http://web.stanford.edu/∼kdevlin/Papers/HPI_SocialSciences.pdf
11. Devlin, K.: A uniform framework for describing and analyzing the modern battlefield, US Army Feasibility Study Report, 19 pp. (July 2011).
    Access: http://web.stanford.edu/∼kdevlin/Papers/Army_report_0711.pdf
12. Englmeier, K. and Murtagh, F.: What can we expect from data scientists?, Journal of Theoretical and Applied Electronic Commerce Research, 12(1), i–iv, (Jan. 2017).
    Online access: http://www.jtaer.com/statistics/download/download.php?co_id=JTA20170100
13. Englmeier, K. and Murtagh, F.: Data Scientist – Manager of the discovery lifecycle. Proceedings of the 6th International Conference on Data Science, Technology and Applications – Volume 1: DATA, 133–140, 2017.
    Access: http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=PZAF7J3pInA=&t=1
14. Darabi, A.: The UK's next census will be its last – here's why, News report, (5 Dec. 2017),
    https://apolitical.co/solution_article/uks-next-census-will-last-heres
15. Escoufier, Y., Fichet, B., Lebart, L., Hayashi, C., Ohsumi, N., Baba, Y.: Editors, Data Science and Its Applications. Academic Press, Tokyo (1995)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

16. Hand, D.: The dangers of not seeing what isn't there: selection bias in statistical modelling, ISA Gosset Lecture 2017, Royal Irish Academy, (7 April 2017)
17. Hand, D.: Statistical challenges of administrative and transaction data, Journal of the Royal Statistical Society, Series A, 181 (3), 1–24 (2018). Including F. Murtagh comments.
18. Hayashi, C.: What is Data Science? Fundamental concepts and a heuristic example, in Hayashi, C., Yajima, K., Bock, H.H., Ohsumi, N., Tanaka, Y., Baba, Y. (eds.). Data Science, Classification, and Related Methods, pp. 40–51, Springer, Heidelberg (1998)
19. Hayes, B.: Empirically-based approach to understanding the structure of data science, business over Broadway, Seattle, Washington. (18 January 2016).
http://businessoverbroadway.com/empirically-based-approach-to-understanding-the-structure-of-data-science
20. Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A.: AAPOR Report on Big Data, Technical Report. AAPOR, American Association for Public Opinion Research, 50 pp., (2015)
http://www.aapor.org/Education-Resources/Reports/Big-Data.aspx
21. Keiding, N. and Louis, T.A.: Perils and potentials of self-selected entry to epidemiological studies and surveys, Journal of the Royal Statistical Society, Series A, 179 (2), 319–376 (2016). Including F. Murtagh comments.
22. Le Roux, B. and Lebaron, F.: Idées-clefs de l'analyse géométrique des données, in Lebaron, F. and Le Roux, B. (eds.). La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données, pp. 3–20, Dunod, Paris (2015)
23. Mahabal, A.A., Crichton, D., Djorgovski, S.G., Law, E., and Hughes, J.S., From sky to earth: Data Science methodology transfer, in M. Brescia, S.G. Djorgovski, E. Feigelson, G. Long and S. Cavuoti, eds., Astroinformatics (IAU S325) (Proceedings of the International Astronomical Union Symposia and Colloquia), pp. 17–26, Cambridge University Press, Cambridge UK (2017).
Article access: https://arxiv.org/pdf/1701.01775.pdf
24. Murtagh, F.: Origins of modern data analysis linked to the beginnings and early development of computer science and information engineering, Electronic Journal for History of Probability and Statistics, 4 (2), pp. 26, (2008). Online access: http://www.jehps.net/Decembre2008/Murtagh.pdf
25. Murtagh, F., Pianosi, M. and Bull, R.: Tracking and mapping Habermas's communicative action: A case study using Twitter social media. Quality and Quantity, 50, 1675–1694 (2016)
26. Murtagh, F.: Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics, Chapman & Hall/CRC Press, Boca Raton, FL (2017)
27. Murtagh, F. and Farid, M.: Contextualizing Geometric Data Analysis and Related Data Analytics: A Virtual Microscope for Big Data Analytics, Journal of Interdisciplinary Methodologies and Issues in Science, Special Issue on Digital Contextualization, Vol. 3, 19 pp., (June 2017).
https://jimis.episciences.org/3936/pdf
28. Murtagh, F., Orlov, M., and Mirkin, B.: Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research, Journal of Classification, 35 (1), 5–28, (2018). Preprint: https://arxiv.org/abs/1607.03200
29. Murtagh, F.: Security and ethics in Big Data: Analytical foundations for surveys, Archives of Data Science (submitted) (2018)
30. NVP, NewVantage Partners, Big Data Executive Survey 2017, Executive Summary of Findings, Big Data Business Impact: Achieving Business Results through Innovation and Disruption, pp. 16 (2017)
Access: http://newvantage.com/wp-content/uploads/2017/01/
Big-Data-Executive-Survey-2017-Executive-Summary.pdf
31. Ohsumi, N.: From data analysis to data science, in Kiers, H.A.L. Rasson, J.-P., Groenen, P.J.F., and Schader, M. Data Analysis, Classification, and Related Methods, Springer, Heidelberg, pp. 329–334 (2000)
32. O'Neill, C.: Weapons of Math Destruction, Crown/Archetype, Danvers, MA (2016)
33. Press, G.: Graduate Programs in Big Data Analytics and Data Science, (Last updated: October 26, 2017)
Access: https://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science
34. Ueno, M.: As the oldest journal of Data Science, Behaviormetrika, 44 (1), 1–2 (2017)
35. United Nations, Committee on the Peaceful Uses of Outer Space, 59th Sesssion, Vienna, 8–17 June 2016: "Open Universe" proposal, an initiative under the auspices of the Committee on the Peaceful Uses of Outer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2018 doi:10.20944/preprints201805.0418.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 14; doi:10.3390/bdcc2020014

Space for expanding availability of and accessibility to open source space science data. Access: http://www.unoosa.org/res/oosadoc/data/documents/2016/aac_1052016crp/aac_1052016crp_6_0_html/AC105_2016_CRP06E.pdf

36. Wessel M.: You don't need Big Data – You need the right data. Harvard Business Review (3 Nov. 2016). Access: https://hbr.org/2016/11/you-dont-need-big-data-you-need-the-right-data
37. Wilkinson, M.D. et al.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3 (2016).