Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What

Peer-reviewed author version

AERTS, Marc; MOLENBERGHS, Geert & THAS, Olivier (2021) Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What. In: ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION, 8 (1) , p. 25 -39.

DOI: 10.1146/annurev-statistics-040620-032820
Handle: http://hdl.handle.net/1942/34112

Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What

Marc Aerts,[1] Geert Molenberghs,[1,2] and Olivier Thas[1,3,4]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium
[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium
[3]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, New South Wales, Australia
[4]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–15

This article's doi:
10.1146/((please add article doi))

Keywords

computer science, curriculum, data science, education, graduate, interdisciplinary, statistics

Abstract

Organizing a graduate program in Statistics and Data Science raises many questions, offers a variety of opportunities while a multitude of choices have to be taken. The call for graduate programs in Statistics and Data Science is overwhelming. How does it align with other (future) study programs at secondary and postsecondary level? What could or should be the natural home for data science in academia? Who meets the entry criteria, and who does not? Which strategic choices play inevitably a prominent role when developing a curriculum? We share our views on the Why, When, Where, Who and What?

Contents

1. Why?

Subsequent waves of innovation to a varying degree have run through the statistics community during the last 75 years: epidemiology and non-linear models such as logistic regression, Bayesian model fitting, complex multivariate and hierarchical models, smoothing methods, bootstrap, data mining, bioinformatics, omics, big data,... Not surprisingly, they were often preceeded by technological breakthroughs in computer science and software development and/or by innovation in one or more substantive sciences. With inter-innovation time lags varying during our history, these waves rippled through all aspects of our work, including in education in statistics, at all levels from secondary school up to postgraduate university programs.

Although the data science wave bears similarities with earlier waves, especially those of the bioinformatics and big data waves, the current "data science" wave is of another, much larger scale. It has not only been initiated by scientific or intellectual developments but especially also by societal, political and commercial developments. More than ever, detailed and personalized data are available (and not always with recognized consent of the individual) and more than ever such data imply knowledge and insights and evidence based decisions and treatment choices to the benefit of the patient or client at best. But other commercial interests may have a dominant impact on how data science will develop further. The current data science hype has many characteristics in common with bioinformatics, including the availability of several big (public and private) data resources and their integration. There is, of course, lack of clarity and even controversy about the definition of data science. One's definition has become in itself a discrete random variable with probabilities highly depending on characteristics of the user's profile (e.g., engineer, computer scientist, theoretical physicist, statistician, etc.). Bioinformatics was, at the early stage, often claimed by computational biologists and/or computer scientists, starting with the flawed linguistic observation that the term bioinformatics itself was an indication that informatics rather than statistics played the centerstage role. Meanwhile, and for over a decade, it has been recognized that statistical methodology is a key component of bioinformatics, and statisticians started to use the term "statistical" bioinformatics to label this component more explicitly. One could think of the data science movement as a final, all encompassing stage in the expanding process that started with bioinformatics in life sciences and then jumped to omics data and actually to big data in any other field of application, and especially to the field of marketing and related business and marketing fields, and further to the social and political sciences.

We do not aim to present yet another definition of data science. Definitions vary from exhaustive lists of subjects of knowledge and skills to just defining data science as the science of learning from data, with everything that this entails (e.g. Donoho 2017), the latter covering knowledge and skills of data management, data wrangling, data visualization, planning of studies, statistical methods, models, and inference, artificial intelligence, machine learning, etc. It naturally also includes sufficient knowledge about the substantive area within which an application or methodological development is situated. Cameras recognizing individual behavior and translating these data into actions, may be perceived as data science.

The wide scientific aura around a data scientist might explain the tendency to rather identify such a data scientist as an engineer, who typically has enjoyed exposure to many scientific fields in an applied fashion during their education, as compared to, for example, a master of statistics taking a primarily mathematical statistics program. We believe that there is room and need for several variations of data scientists, varying in their particular knowledge and skill sets but having sufficiently overlapping subsets. All are needed in our complex contemporary society.

In Section 2 we discus at which stage in an individual's education what exposure should be given to data science. Where such exposure and education should be administered is the topic of Section 3. Who holds responsibility over it and should be involved is tackled in Section 4 and what a curriculum in data science at graduate level should look like is approached in Section 5. An outlook to the future is offered in Section 6.

2. When?

Already at young ages, children can observe, discover, and experience basic concepts of data science in their everyday environment with quantitative and qualitative data being omnipresent in our society, with basic practice in collecting and sharing data, basic use of information and evidence from data, including sensitivity of certain data, etc. But little seems to be available on educational strategies implementing a systematic progressing of data science knowledge and skills across curricula, covering the age range of, say, 5–18 years of age. Heinemann et al. (2018) developed a data science curriculum for German secondary schools, designed as an interdisciplinary approach between mathematics and computer science education, with also a strong focus on societal aspects. An interesting and inspiring report entitled "The integration of data science in the primary and secondary curriculum," addressed to the Royal Society Advisory Committee on Mathematics Education (Pittard 2018), reviews Britain's primary and secondary national curricula (Key Stages 1–4), identifies the extent to which elements of data science exist within data-rich subjects, and identifies specific topics that are conducive to nurturing data science skills. Main challenges, barriers and strategies to overcome these, as well as further opportunities are discussed as well. We can only recommend other governmental authorities worldwide to conduct similar insightful exercises.

Its interdisciplinary nature, more extensive and complex prerequistes, numerous application areas, rapid scientific development, huge demand for data scientists in industry, and growing impact on future lives make education in data science a very demanding but challenging field, for students, parents, teachers, curriculum developers, schools, universities, scientific societies, and governments. It is unquestionable one of the grand challenges of education, contemporary as well as in the near future. It also offers unique opportunities. These challenges cannot be faced and opportunities cannot be taken by individual teachers or single organizations or authorities only, and the scientific and educational communities can hope for, but not wait for appropriate measures and guidelines from governmental bodies, as these will be, in all likelihood, ill-timed, ill-conditioned, and with too limited an outlook. All initiatives are very much welcomed, but as data science is intrinsically interdisciplinary, the approach for data science education to face the challenges and to grasp opportunities should be "inter-" in many facets as well: interdisciplinary, inter-curricular, inter-grade, . . . , around inter-collaborative initiatives across study programs of different age groups, with involvement of secondary school teachers and professors, etc. Statistics education at primary and secondary schools has been facing a deficiency in skilled statistics teachers since it got a more prominent place in the curricula, and this will not be different for data science, on the contrary. Actually, in many countries, this capacity problem is extending to other disciplines and the number of teachers with the appropriate diploma for teaching particular courses decreases gradually. For the rapidly evolving field of data science, even not having its own place and list of end competences in the primary and secondary curricula, a solution could be to support and facilitate collaboration between schools and universities. Here universities can play a leading role, next to offering bachelor and master programs in statistics and data science. Such support could come from grants in innovative educational projects at universities, government, industry, etc. Similar to or part of the STEM initiative (Science, Technology, Engineering and Mathematics, see e.g. Hallinen 2019), there is need for an education policy and curricular choices in schools to educate students in data science. A diverse palette of organized learning activities could stimulate and excite youngsters' interest in data science: professors acting in "meet the professor" events in secondary schools, senior students taking a master of statistics and data science could be involved in teaching data science in secondary school (and getting credit for it in their own study program), student-driven science fair projects and citizen science (see e.g. Koomen et al. 2018, for an interesting account of middle school science fair projects inspired by citizen science monitoring), . . . Of course, this should not just be organized with fragmented and isolated learning units, but rather fit in a strategic plan with a grow-curve pattern for interested students.

Turning to undergraduate programs in statistics and data science, the 2015 special issue 69:4 of *The*

American Statistician on Statistics and the Undergraduate Curriculum includes several contributions on data science in statistics curricula, providing several examples and resources for instructors to implement data science in their own statistics curricula. De Veaux et al. (2017) proposed detailed guidelines for revising a major in data science, starting with redesigning existing courses in computer science, statistics and mathematics, and guided by the relation of data science to the other sciences. In the next stage, they suggest to transition existing courses further and develop new courses to more fully integrated courses, taking advantage of the efficiencies and synergies that an integrated approach to data science would provide. They propose a case-based and hands-on approach, as is common in fields such as "engineering and computer science." We would like to add the field of applied statistics to this. We can only fully agree with the statement that the two pillars, computational (algorithmic, predictive) and statistical (inferential) thinking, should not be kept separately but rather integrated. Yavuz and Ward (2020) discuss the implementation of an introductory data science course, offered during a student's sophomore undergraduate training in statistics. It implements many of the best practices espoused by De Veaux et al. (2017), including interaction and a strong focus on teamwork.

De Veaux et al. (2017) mentioned in their report that the website http://datascience.community/colleges listed at that time 530 programs in data science, analytics, and related fields at over 200 universities around the world. At the time of writing, the web site lists already 618 programs, with still a vast majority of these being masters degree and certificate programs offered both traditionally and online.

Since December 2016, the U.S. National Academy of Sciences (NAS) has been holding a series of roundtables on data science postsecondary education. Each of four meetings per year focuses on a topic related to data science education or practice. These roundtables bring together representatives from academic data science programs, funding agencies, professional societies, foundations, and industry to discuss the community's needs, best practices, and ways to move forward. The objective is to "help affected communities develop a coherent and shared view of the emerging field of data science and of how best to prepare large numbers of professionals to help realize the potential of this field." The written summaries are available at `https://nas.edu/dsert` and are a rich source of information, based on the views of several experts with different backgrounds, experiences and interests. Topics include the development of data science curricula and programs at 2-year colleges and its importance in the development of a diverse and inclusive workforce (one-third of the total undergraduate student population in the US is enrolled at 2-year colleges, citing Nicholas Horton from the minutes of the meeting #11); promotion of data science for socially desirable outcomes; content, organization and alternative structures of data science PhD programs, and many more. At the European side, there is the European Data Science Academy (EDSA) funded by the Horizon 2020 Framework Programme of the European Union (`http://edsa-project.eu/`). Mikronyannidis et al. (2018) present this European initiative for bridging the data science skills gap across Europe and training a new generation of world-leading data scientists. The EDSA project has established a rigorous process and a set of best practices for the production and delivery of curricula and courseware for data science, as well as linking demand with supply.

So, on the question when particular data sciences learning units on statistics and data science should be offered, we would like to respond: "at any age, starting at early age, and continuing lifelong." Not everyone will agree with this point of view. One could, for example, argue that it is better to wait until a student has been confronted with the urgency to learn statistics and data science when trying to analyse their own data (e.g., from a bachelor thesis). This genuine interest in data however can also be obtained for children in secondary school, by connecting with daily real-life situations and active citizenship. For instance, the COVID-19 crisis (2020) offers endless opportunities to connect with data, data on social contacts illustrating the importance of measures to change human contact behaviour, numbers of infected, hospitalised and deceased individuals, and to discuss sensitive issues around such data, to compare curves

and discuss time trends shown daily on the internet, etc.

Also, continuing education and lifelong learning become more important, as the pace of scientific and technological innovation keeps on accelerating. Education in statistics and data science needs a holistic approach, maintaining and upgrading knowledge and skills in expanding contexts and environments. Therefore it is important to learn students at bachelor and master level skills for self-learning, self-motivation, self-monitoring, self-adjustment, self-sustainability, ... skills on which we will come back in Section 5.

In our view, a full 2-year study program of say 120 study credits on statistics and data science, is ideally offered as a master program. We also think that initiatives across different study program and age groups, triggering the general interest of students at young age, stimulating collaboration across any border, and filling in gaps of capacity of expertise in statistics and data science at particular levels (e.g., by master and PhD students, postdoctoral researchers, and professors in the field of statistics and data science contributing to the education at secondary level) can only be welcomed and should be supported in various ways. Industry can and should also get and take the opportunity to play its role (e.g., data scientists in industry acting guest professors in master programs or external supervisors of master theses, or supporting and participating in so-called hack weeks (Huppenkothen et al. 2018), etc. Joined forces are highly needed to answer effectively the high and growing demand for statisticians and data scientists coming years. Whereas statistics has been branded too often as "boring", data science (being the sexiest job of the 21 century, Davenport and Patil 2012) enjoys the interest of youngster much more. The momentum is there and passing by . . . ; all data scientists should consider contributing to education in data science.

3. Where?

Hicks and Irizarry (2018) argue that a statistics department that embraces applied statistics is a natural home for data science in academia. While statistics is not the only field contributing to data science, and while arguably there is no uniform best organizational format, there are several reasons why the statistics discipline might beneficially be placed at the center of organizing data science. Almost invariably, successful applied statistics groups, whether focusing on a single or various substantive areas, bring under a single roof research, teaching, and consulting. When both methodological research as well as collaborative research with substantive fields is undertaken, such a department is in a good position to ensure sufficient breadth and depth of the data science curriculum. The focus in research and consulting that a group has should ideally translate in the signature features of the curriculum offered. For example, a group focusing primarily on biostatistics may be well-suited to lead a graduate program in data science for biostatistics, but it may be challenging to support a program that has other substantive pillars, such as engineering or economy. This is important for master level education, and even more so for PhD level training, to ensure that PhD candidates are able to work on relevant and challenging projects of sufficient breadth.

This suggests that there should be close collaborations between the statisticians and data scientists on the one hand, and the substantive areas covered in the graduate program on the other hand. Hicks and Irizarry (2018) argue that applications should be brought much more to the forefront than is currently the case. The problem should come first and the teaching should be linked to one or a number of substantive problems, even when more theoretical and abstract concepts are being taught. It is possible that a given institution does not encompass all the specialties envisaged. Wherever possible, it is advisable to consider inter-institutional collaborative efforts. This is relatively easy in geographical areas where several complementary institutions are available at relatively short distance. For example, a university with a large engineering school could collaborate with an institution that is known for its medical research. Fortunately, because of the increasing availability of electronic and distance learning tools, the physical distance between institutions is becoming less and less of a limiting factor.

The above makes clear that a signature feature of applied statisticians is the habit to collaborate

within their own field with colleagues with different substantive interests on the one hand, and with colleagues from the substantive fields on the other. This is a vital asset because data science requires more, not less, collaboration, to give curricula the broad and high-quality basis they deserve. In other words, the collaborative network is expanding thanks to data science, with in particular engineering, numerical mathematics, and computer science groups being added, as well as additional substantive teams, bring about more so-called use cases.

In some institutions, often the larger or older ones, there may not be a single applied statistics group, but various statistics groups may co-exist, allowing that each group has a symbiotic relationship with the substantive area of their specialty. In others, often smaller or younger ones, statisticians working with various substantive areas may find themselves in the same organizational unit. In both cases, it is perfectly possible to organize a data science graduate curriculum. Hasselt University in Belgium is a smaller, younger university with a relatively large Center for Statistics. Their research, consulting, and graduate teaching all focus on biostatistics, epidemiology and public health methodology, and bioinformatics, which are then quite naturally the themes within the Master of Statistics and Data Science. A Data Science Institute encompasses this structure, as well as researchers from computer science, medicine, and economy. Data Science Institutes that take the lead in a data science program exist at Columbia University (encompassing computer science, statistics, engineering, and operations research) and at the University of Virginia (with computer science, statistics, and systems engineering participating). At the large and older Leuven University, statisticians are dotted around over ten different faculties, but their paths cross in the university-wide Leuven Statistics Research Centre, which in turn is the organizational structure for the Master of Statistics and Data Science. So indeed, when statisticians are spread over various entities, it is important that there are efficient communication lines between them and, ideally, that there is a superstructure, such as a university wide center or institute for statistics (and data science). At Harvard University, the Institute for Applied Computational Sciences groups statistics, computer science, and applied mathematics. In various places, a data science program is organized jointly by various quantitative departments, without there being a superstructure: at Kennesaw State University, statistics, computer science, and mathematics all contribute; at the University of Colorado, this is biostatistics and informatics, and computer science and engineering; biostatistics and medical informatics are the contributing fields at the University of Wisconsin-Madison; at the University of British Columbia, statistics and computer science co-organize the program; Tufts University sees a collaboration between computer science and engineering as the basis for their data science master, i.e., an example where statistics is not formally part of the organizational structure. At the University of Kansas, a data science program is organized by the biostatistics department, and to this end statisticians with computational background are recruited to the faculty. At the University of Toronto, the Master of Applied Computing is led by the computer science department, with the data science concentration in the hands of the statistics department. In various other places, there is a partnership between statistics and a substantive area. For example, collaborations between statistics and mathematics on the one hand, and the business or management school on the other, exist at Texas A&M, Clemson University, and Massachusetts Institute of Technology.

This underscores that there is no single best way of organizing data science and a corresponding graduate education program. But the institutional environment should ensure that the recursive data cycle (De Veaux et al. 2017), made up of obtaining, wrangling, curating, managing, and processing data, is steeped in practice. Likewise, Wild and Pfannkuch (1999) referred to this as the "problem, plan, data, analysis, conclusion" cycle. No matter what the institutional context is, the more researchers from various disciplines are integrated and feel to be part of a larger entity, the higher the chance that the cycle will be effective.

Because of the interdisciplinary and multi-disciplinary nature and the extent of the collaborative network needed, it is futile to assume that establishing a single data science department, rigidly existing next

to other departments, is a fruitful approach. Rather, because of the multi-disciplinary nature, efficient networking-type structures are needed, of which academic personnel can be a member without having to give up the membership of their conventional department. Kane (2014) states that statisticians are situated at the confluence of statistics, computer science, and the substantive area. Cleveland (2014) itemizes the various disciplines and fields that should be present in data science research, which mirrors the needs in the accompanying graduate programs, of course. They are: multidisciplinary investigation, models and methods for data, computing with data, pedagogy, tool evaluation, and theoretical foundations of data science.

Hicks and Irizarry (2018) refer to type A and type B data scientists. Type A is concerned more with the statistical and methodological side, in view of answering real-world questions; Type B refers to the coding side of data science, and is rooted in engineering and computer science. In a way, the aforementioned Tufts University example is of Type B. Whereas Hicks and Irizarry (2018) focus in their guide to teaching data science on Type A, arguably at graduate level it is important that both types are brought together in a flexible, perhaps virtual structure that facilitates interaction and collaboration.

Evolutions of this type are not new to statisticians, even though it may seem so in the current day and age. One of the earlier similar events was the advent of epidemiology. To this day, there is a variety of ways in which statisticians and epidemiologists have organized themselves around each other. It is telling that in many institutions scholars belong to both. Similar evolutions took place with the coming of, respectively, statistical genetics, bioinformatics, the omics, and big data.

Broadly, the drivers of new evolutions are either a quantum leap in a substantive field (e.g., the decoding of the genome), necessitating new or expanded quantitative methodology, a major step forward in quantitative sciences (e.g., increasing computational power; data capture from social media), or both (wearable medical devices). One often drives the other. For example, the advent of generalized linear models, such as logistic regression, very commonly used in epidemiology, was turned into practical use thanks to advances in computer hard- and software. In this sense, Cleveland's (2014) broad view on data science is important: "A very limited view of data science is that it is practiced by statisticians. The wide view is that data science is practiced by statisticians and subject matter analysts alike, blurring exactly who is and who is not a statistician."

We can learn from these earlier (r)evolutions in many ways. For example, programs in epidemiology and public health may be stand-alone or organized jointly with (bio)statistics programs. The same is true for bioinformatics. Various organizational forms are possible and also here there was no single best format. For example, at Hasselt University, the Master of Statistics and Data Science encompasses tracks in biostatistics, quantitative epidemiology, bioinformatics, and data science. At KU Leuven, the Master of Statistics and Data Science mainstreams data science across a variety of tracks, that are substantively oriented (biometrics; social, behavioral and educational sciences; industrial applications; business; official statistics; and theory). On the other hand, the bioinformatics curriculum at KU Leuven is stand-alone, with some courses in common with the master of statistics and interchange of faculty.

It seems evident, in this day and each, that we need to make our programs resilient. So, they should not purely rely on imparting knowledge in a face-to-face way. Rather, they should be interactive and involve two-way traffic between students and instructors. The use of distance-learning and flipped-classroom techniques should be part of a program in a routine fashion. Like with the 2020 COVID-19 crisis, it is of the essence that a program can revert to a fully on-line version by merely "flipping a switch". High quality programs may make use of MOOCS, or at least SPOCS, to this effect.

4. Who?

Flexibility is needed to accommodate various types of prior education, at the undergraduate level but, for some students, also in terms of their prior graduate education. It is now common that many students choose graduate education, or further graduation education, in institutions different from the ones they got earlier degrees from. Careful assessment of their prior knowledge and skills is necessary, not only to decide on admission, but also to gauge whether a reduced and/or tailor-made curriculum is advisable. We can think of three types of students admitted to a data science graduate program: (a) those with undergraduate training in data science; (b) those with undergraduate training in one or a few of the contributing fields, i.e., statistics, engineering, computer science, or mathematics; (c) those with undergraduate training in a sufficiently quantitative subject area. All of this implies that the student population will be heterogeneous, which is challenging, but also foreshadows the working environment that most graduates will end up in. Relatedly, it offers opportunities for group assignments in multidisciplinary teams.

When (transition to) new or extended programs in statistics and data science are discussed, a lot of attention typically goes to curriculum guidelines but less to the admission process. However, every academic year again, Admission and Examination Boards experience the importance of an efficient and well calibrated admission process. Setting the entry conditions on the required prerequisite knowledge too low implies that too many students will fail and hence will not earn the degree, for some of them after three to four years of study. This has to be avoided for many reasons, the most important being the future of the student concerned, next to cost-efficiency considerations. Setting it too high is withholding the unique opportunity for realizing the dream to become a professional statistician/data scientist for those who do not meet the entry conditions but do have the potential to fill in the gaps and make required progress en route.

The starting point is an unambiguous set of verifiable entry competences, a clear communication about it to candidate students and an admission process built upon verifiable information and valid proof that the candidate meets the entry criteria. In our view an applied master in data science program should be open not only to bachelors in statistics, mathematics, and engineering, for example, but also to bachelors (and masters) in biology, life sciences, medicine, economy, chemistry, sociology, psychology,... More generally, bachelors and masters coming from fields where data are the basis of knowledge, evidence-based decisions are taken, and having sufficient quantitative knowledge and potential for a sufficiently high level of abstract thinking, can be considered for admission. Evidently, more so than ever, individual assessment before entry will be needed, which is a challenging but necessary endeavor. Evidently, in the transition from a statistics to a data science program, the admission process needs to transition along with it.

5. What?

Hicks and Irizarry (2018) started their paper with raising the question: "What is missing in the current statistics curriculum?". This is a very natural question for statistics programs that intent to widen their scope towards data science. An important perspective on this question comes from the job market. Zheng (2017) argues that classically trained statisticians nowadays find themselves frustrated and disappointed when they realize that their competitors from computational sciences perform better in job-interviews or hack-a-thons involving heavy computational skills. More and more job announcements explicitly ask for data scientists, rather than statisticians.

The previous paragraph started from the assumption that a data science program is built on an existing statistics program. However, there are other strategies: one could start from scratch or from another existing program, such as computer science, engineering, mathematics, etc. In this paper, we only discuss how a data science curriculum can be added to a statistics program. It is important to keep in mind an important implication in that graduates from such programs should be data scientists and statisticians

simultaneously. This will have consequences that will be discussed later.

We will make a distinction between two types of statistics programs. A program may have a horizontal structure, which indicates that tracks in certain specializations are offered in a modular system. Students choose modules so as to orient their studies towards a specialization (e.g., biostatistics, business statistics, social statistics, bioinformatics). Thus each student may compose their own individual program (obviously with some restrictions applying). In a vertical program, on the other hand, all specializations have many courses in common (e.g., in the first year), and the specialization-specific courses are mostly offered in the second year. This structure typically leaves only limited room for elective courses.

When a horizontal program is changed into a Statistics and Data Science program, the new data science courses are typically common for all students. This ensures that all students can have the "data science label" upon graduating. If the data science courses would only be offered as a module, not all students would necessarily acquire the data science skills. For statistics programs that extend their scope in this way, the names of the tracks could then be changed from, for example, "Biostatistics" to "Biostatistics and Data Science," i.e., adding the Data Science label to all tracks.

In a vertical structure, on the other hand, new data science subjects can be added to the common part (mostly in year 1) and a new specialization "Data Science" can be created. In this way, all students will have a firm basis of statistics and data science, but students in the Data Science specialization can deepen their data science knowledge and skills. In Belgium the new programs of Master of Statistics and Data Science at KU Leuven and Hasselt University are examples of a horizontal and vertical structured curriculum, respectively. Despite this fundamental difference, both programs share a few courses.

Before addressing what data science topics can be included in a curriculum, we want to state the obvious truth that adding more courses to a curriculum, necessarily implies the reduction of content among the existing courses. Thus, when changing a traditional statistics program to a statistics and data science program, quite a few statistical topics need to be sacrificed, courses merged, and content pruned from redundant or overly elaborate topics. The search for these topics is perhaps as important as the selection of new data science topics to be added. The challenge of adding new and deleting existing content could or should go closely together with the challenge of applying innovative teaching methods, optimizing learning efficiency and controlling course/study load. Blended or hybrid learning combining online with traditional classroom methods (e.g. flipped class rooms) tailored to the learning activities within or across courses, exposes and enriches students with important skills such as self-learning, self-monitoring and self-adjusting, being key-skills for lifelong learning.

Hicks and Irizarry (2018) argue that the missing topics in current statistics curricula are related to: computing, connecting, and creating. "Computing" refers to programming skills and general knowledge about computing infrastructures and organization (parallel computing, distributed computing, data bases,...). Data scientists should be able to "connect" their skills to subject matter questions. With "creating" they mean that the data scientist should take part in the creative scientific process and that he/she even must formulate new substantive research questions. They argue that courses should be build around diverse case studies, include computing in almost every aspect of the course and minimise the use of mathematical notation. They continue by recommending that all course activities should mimic a realistic data scientist's experience. This is in contrast with typical courses in a conventional statistics curriculum where students get almost always well-defined homework assignments that start from clean datasets that are provided to them in a format that can be directly read by the software. This observation was also made by Zheng (2017), who concludes that statistics students do not get enough experience with the entire data analysis cycle. He also suggests to expose students much more the realistic situations with messy data. A similar data-centric approach has also been put forward by Hardin et al. (2015). From this literature (and the references therein), there seems to be a broad consensus for curricula with many data-centric courses that integrate computing skills, algorithmic thinking, and, of course, statistical thinking.

Although we agree with many of these suggestions, we do have a few remarks. Programs that aim at broadening their scope from statistics to statistics and data science, either horizontally or vertically, do not necessarily need to redesign all their statistics courses to make them data-centric or full of computational aspects. We believe that there also must be data scientists that have the skills to read and understand the large body of statistical literature. This requires a good understanding of the traditional statistical theories and conventions, because they still stand and they form a strong basis for many data analyses. If graduates from data science programs would no longer be capable of reading and understanding the vast statistical literature, then this knowledge would be lost to them. Moreover, we see advantages in deliberately integrating course activities for which no computers may be used. Thinking and reflecting on problems and solutions are also important skills; the current generation of students is often distracted when sitting in front of a computer. Also for the design of experiments and studies, computers are not required in all phases. Another motivation for not redesigning all courses is more of a practical nature: sometimes courses are shared with other programs. For example, a programming course (e.g., Python) may be offered to student groups from different master programs (e.g., engineering, mathematics, physics, ...). The organization of a separate course for only the statistics and data science students may require too many resources. In such cases, as a compromise solution, the homework or project assignments may be differentiated between the student groups, or a separate tutor for PC-labs may be assigned for the statistics and data science students. These adaptations come at a lower cost and may still help in targeting these shared courses to a data science audience.

Coming back to the position of mathematics in the program, we refer to De Veaux et al. (2017), in which the recommended curriculum refers to mathematical foundations and statistical modeling. In fact, we can refer to this as the required mathematical foundation of statistics (i.e., probability and statistical inference) on the one hand and statistical techniques on the other. At the graduate level, it is important to connect to the knowledge and skills already acquired during prior education and to build upon these to deepen and refine the knowledge and skill sets. The data scientist should dispose of a sufficiently large toolkit of statistical models and other techniques, together with a clear understanding of their uses, advantages and limitations from a mathematical and computational perspective. They should be able to examine the properties of newly proposed methodology, including by themselves. This can but need not be based on mathematical arguments; there is also room for simulations, for example. So, even profound theoretical concepts should not be taught in a context-free way, but have to be linked, as generically as possible, to the statistical and data science practice. This connects to the fact that the successful data scientist should have the habit of following up new evolutions in any of the underpinning disciplines, not just to know and use whatever is new, but also to assess merits and pitfalls in a systematic, critical way.

Earlier we discussed the suggestion to make all courses data-centric and computing-oriented, and design all courses around case-studies and make the students start from extracting or scraping messy data from databases or the internet. Another curricular element that may be considered to reach more or less the same goal, is a capstone project. For example, in each semester one capstone project may be part of the curriculum. It may be organized as a full week, in which no other lectures are scheduled. To some extent it can be compared to a hack-week, which was also proposed as a model for data science education by Huppenkothen et al. (2018). Students may work in small groups on a large real-world case study, and the methods and skills required to solve the problems should come from the other courses taught in that same semester. Lecturers of those other courses may voluntarily enter the PC-lab in this project week and may help or discuss issues with the students, or they may even do some just-in-time-teaching when appropriate.

Whether or not all courses become data-centric and make students go through the whole data analysis cycle, we also believe that there should be better connections between the courses as compared to many traditional statistics programs today, in the sense that at least some of the fundamental data management and programming skills should return in many courses. When, for instance, SQL is only used in the

data management course, students may be unclear as to why they have to acquire knowledge about this language.

Thus far, we focused mainly on adding more computational skills to the curriculum, as well as bringing and keeping the students closer to the data. With respect to the former, there seems to be a consensus today that both R and Python have their place in a statistics and data science program. Many new programs offer a course about data management, which may cover topics related to databases, querying, data representations, etc. Data wrangling may be part of such a course, or it may be part of another course (e.g., basic programming in R or Python). Data wrangling or data cleaning is important to prepare students for dealing with the whole data analysis cycle. We mentioned already the importance of algorithmic thinking for data scientists. It is important to first decide whether a curriculum wants algorithmic *skills* or algorithmic *thinking* as a learning outcome. If it is the former, then a separate course, or sufficient attention to algorithms in a programming course, may suffice. However, when algorithmic *thinking* is the objective, it should be integrated in many courses. Combinations of both approaches are of course also welcome.

Another consensus topic is data visualization. When data comes to us as simple matrices, then most scientists know how to visually explore the data, even for high-dimensional data many (new) visualization tools exists. It becomes less obvious when data can no longer be represented as a matrix or when multiple data sources or data types need to be combined for visual exploration (e.g., networks, audio, video, ...). Because of these complexities, there is a need for a separate course on this topic, or it should be sufficiently covered in courses on data management, data visualization and computing.

An important part of the curriculum should be devoted to awareness and active pursuit of good scientific practices, ethical behavior, and protection of privacy (Keller et al. 2016). The core curriculum should contain goals and methods for reproducible research, including version control, well-annotated software, seamless linking of data, analysis and reporting. The use of packages such as R markdown, Sweave, Overleaf, Jupyter notebooks ... and getting familiar with software development platforms such as Github are typically part of courses on programming in R or Python and data management, but such tools should be used consistently in different learning units or courses across the whole study programme.

Saltz et al. (2018) identified twelve key ethics themes that should be present in a data science curriculum. They conclude that no single existing code of conduct or ethics framework covers all issues and so there is need for more research in this rather new area. They also make the point that apart from general ethical values that should hold for all data scientists, ethics highly depend on the context and the field of application. For example, working with sports data requires another ethical framework than working with medical data. Although one may consider adding a course devoted to the ethics of data science, we believe that ethics should be touched upon in most of the other courses. In this way, the ethical considerations can be discussed in particular contexts (e.g. informed consent in a clinical trials course, ...). In a similar fashion, legal aspects (e.g., privacy / GDPR) can be touched upon in courses when relevant. Indeed, the vast availability of so-called non-designed data (these authors estimated it to represent about 70% of the available data, made up of digital data and social media exchanges), requires careful reflection on privacy and corresponding measures to ensure confidentiality. It is interesting to see that time honored techniques, such as multiple imputation (Rubin, 1987), can be used to ensure confidentiality, as noted by Keller et al. (2016). Multiple imputation can preserve important structure in the data, while avoiding that the data being analyzed can reveal peoples identity. Moreover, the statistical properties of such data, often poorly understood, should be given careful study, quite opposite to the common belief that the unbelievably vast amounts of data available obviate the need to worry about statistical pitfalls (e.g., bias).

The qualified data scientist should carefully understand the concept of differential privacy (see e.g. Dwork and Roth 2014), in the sense that aggregate information (e.g., demographic information) or patterns in data (e.g., contact tracing in the context of an epidemic) are available, perhaps even publicly so, whilst

protecting the privacy of the individuals on which the data is based. In addition to understanding, the ethical need for it should be internalized, and the advantages but also pitfalls carefully understood. In other words, it should become a natural habit to think in terms of differential privacy, whenever applicable.

Perhaps privacy may sound non-committal, because at first it may seem like immaterial to the core content of these courses, or perhaps even to the lecturer. This concern may also hold for ethics and also to other competences that at first sight are far from the core topic of a course. Even algorithmic thinking and computational skills may not sound evident to all lecturers who have to change their course from a statistics to a statistics and data science program.

It is thus important that every program starts from a well formulated vision on the education of data scientists, and that this vision is communicated, understood and supported by all lecturers. This vision should be translated into competences and skills, which should be further made concrete as learning lines. It is the responsibility of the program committee to ensure a logical structure among the topics, arrange them into learning lines and to communicate to the lectures what topics should be incorporated in what course. This is particularly important for ethics and legal aspects, as these do usually not belong to the core technical content of a course.

We have discussed ethical and legal aspects relevant to data science, but there are other skills and competences that are equally important, but not very specific to data science. For example, soft skills such as communication and employability skills (e.g., stakeholder awareness, self management, ...). Many of these are not specific to data science, but they may need to be redesigned to match better the profile of a broad data scientists.

Earlier we mentioned the role of a program committee as central to developing the vision of the program and the formulation and coordination of the learning lines so as to guarantee that all competences are covered at the right place and time in the curriculum. When a traditional statistics program is moving towards a statistics and data science program, it seems important to us that the program committee gets broadened as well, by attracting lecturers of the non-statistical data science subjects (e.g., computer scientists).

Coming back to the relationship between data science and statistics, which is central to the development of a program, we like to stimulate thinking about the added value of statistical thinking and the statistical tradition.

In the development of a statistics and data science curriculum, one should also carefully think about connecting the knowledge, skills and competences the students have acquired in their undergraduate studies with the master level courses. For example, as indicated in De Veaux et al. (2017), careful attention should be devoted to data preparation and data management, to deepen skills acquired at undergraduate level or, if they have not been acquired yet, to mend this.

Models and techniques studied at undergraduate level, should be taken to the next level. For example, models at undergraduate level are typically of a univariate or classical multivariate nature. However, there are many and potentially rich but complex correlated data structures that should be recognized and properly handled. In the same vein, a coherent treatment of missing data problems and solutions should be included. Especially with very big data collections, it is tempting to believe that there are "enough data to overcome the missing data problem," which is of course a fallacy. The risk of bias resulting from incomplete data, from confounding, and from selective sampling mechanisms should be understood and strategies should be offered to handle these well.

Unlike at undergraduate level, graduate education can benefit from a clear focus on one or a few substantive sciences. Evidently, this need not be a uniform choice for a given curriculum, but there can be several tracks to cater to various interests and student subpopulations. Arguably, basic courses in the substantive science should be given, not with the aim to make the data scientist a specialist in the substantive area, but to give him the language, habit, and comfort to successfully interact with specialists

in the area.

An interesting perspective is given by Hernán et al. (2019), who say that building a new data science curriculum should be taken as an opportunity to leave the conventional approach to statistical teaching behind, not only to include more computational and algorithmic competences, but also to bring causal inference into the data science programs. This view may be considered as too extreme and a good balance should be struck across competences.

6. Wrapping up and looking ahead

Many publications on the topic of teaching data science emphasize the need of structuring the course activities to realistically mimic a data scientist's experience. We hold the opinion that several current applied master of statistics programs offer such course activities and could thus be extended to a data science program, but we argue that there should be a good balance of different type of courses. Not all courses should be organized in this manner, as such activities often inevitably imply several concepts and knowledge to be fragmented and scattered and not discussed within a framework that offers sufficient background and insights. Therefore, there is, in our view, across parts of one single course, or preferably across courses, need for applied experience courses as well as more focused courses on a particular topic, being e.g. a course on longitudinal data models or an advanced programming course. So, richness in type of courses and experiences for the student are in our view key to obtain a good balance in methodological insights, key competences and skills etc.

The issue of ethics, privacy and data protection and related regulation should get sufficient attention in the curriculum. Passive knowledge of and awareness about is essential, but it is to be highly recommended that students have an active experience on these key issues, being part of a consulting class in a project course or within the framework of the master thesis. Elliott et al. (2018) call attention to the importance of teaching cross-cultural ethics. Many statistics and data science study programs have a multicultural student population, including Western, Eastern, African cultures. Cultural but also linguistic differences form barriers for an ethical decision-making compatible across philosophical and theological models.

ACKNOWLEDGMENTS

The authors thank the editors and the reviewers for their constructive comments and suggestions.

LITERATURE CITED

Cleveland WS. 2014. Data science: An action plan for expanding the technical areas of the field of statistics. *Statistical Analysis and Data Mining*, 7: 414-417. doi:10.1002/sam.11239

Davenport TH, Patil DJ. 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90, no. 10 (October 2012): 7076.

Donoho D. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

DeVeaux RD, Agarwal M, Averett M, Baumer B, Bray A, Bressoud T, Bryant L, Cheng L, Francis A, Gould R, Kim AY, Kretchmar M, Lu Q, Moskol A, Nolan D, Pelayo R, Raleigh S, Sethi RJ, Sondjaja M, Tiruviluamala N, Uhlig P, Washington T, Wesley C, White D, Ye P. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4(1):15–30.

Dwork C, and Roth A. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9: 211–407. DOI: https://doi.org/10.1561/0400000042.

Elliott AC, Stokes L, Cao J. 2017. Teaching ethics in a statistics curriculum with a cros-cultural emphasis. *The American Statistician*, 72(4):359–367.

Hallinen J. 2019. STEM Education Curriculum. *Encyclopedia Britannica.* https://www.britannica.com/topic/STEM-education

Heinemann B, Opel S, Budde L, Schulte C, Frischemeier D, Biehler R, Podworny S, Wassong T. 2018. Drafting a data science curriculum for secondary schools. *ACM International Conference Proceeding Series* 1-5. DOI: 10.1145/3279720.3279737.

Hicks SC, Irizarry RA. 2018. A Guide to Teaching Data Science. *The American Statistician*, 72(4):382–391.

Hardin J, Hoerl R, Horton N, Nolan D, Baumer B, Hall-Holt O, Murrell P, Peng R, Roback P, Temple Lang D, others. 2015. Data science in statistics curricula: Preparing students to "think with data". *The American Statistician*, 69(4):343–353.

Hernán MA, Hsu J, Healy B. 2019. A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1): 42–49.

Huppenkothen D, Arendt A, Hogg DW, Ram K, VanderPlas J, Rokem A. 2018. Hack weeks as a model for data science education and collaboration. *Proceedings of the National Academy of Sciences of the United States*, Sept 4, Vol.115(36), p.8872(6).

Kane MJ. 2014. Commentary: Clevelands Action Plan and the Development of Data Science over the Last 12 Years. *Statistical Analysis and Data Mining*. doi: 10.1002/sam.11244

Keller SA, Shipp S, Schroeder A. 2016. Does Big Data Change the Privacy Landscape? A Review of the Issues. *Annual Review of Statistics and Its Application*, 3(1):161–180.

Koomen MH, Rodriguez E, Hoffman A, Petersen C, Oberhauser K. 2018. Authentic Science with Citizen Science and Student-driven Science Fair Projects. *Science Education*,102(3): 593-644.

Mikroyannidis A, Domingue J, Phethean C, Beeston G, and Simperl E. 2018. Designing and delivering a curriculum for data science education across Europe. *Advances in Intelligent Systems and Computing* 716, (M. E. Auer et al. eds., Teaching and Learning in a Digital World), https://doi.org/10.1007/978-3-319-73204-6 59.

Pittard V. 2018. The integration of data science in the primary and secondary curriculum. https://royalsociety.org/topics-policy/publications/2018/integration-data-science-in-primary-secondary-curriculum/.

Saltz JS, Dewar NI, Heckman R. 2018. Key concepts for a data science ethics curriculum. SIGCSE'18: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, 952–957, https://doi.org/10.1145/3159450.3159483.

Suman AB, Pierce R. 2018. Challenges for citizen science and the EU open science agenda under the GDPR. *European Data Protection Law Review*, 4(3):284 – 295.

Wild CJ, Pfannkuch M. 1999. Statistical thinking in empirical enquiry. *International Statistical Review*, 67, 223–265.

Yavuz FG, Ward MD. 2020. Fostering Undergraduate Data Science. *The American Statistician*, 74:1, 8-16, DOI: 10.1080/00031305.2017.1407360

Zheng T. 2017. Teaching data science in a statistical curriculum: Can we teach more by teaching less? *Journal of Computational and Graphical Statistics*, 26(4):772–774.