Visualization for Data Scientists: How specific is it?

Beatriz Sousa Santos[1] and Adam Perer[2]

[1]DETI/ IEETA- Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Portugal
[2] Human-Computer Interaction Institute, Carnegie Mellon University, USA

Abstract

Data Science has been widely used to support activities in diverse domains as Science, Health, Business, and Sports, to name just a few. Theory and practice have been evolving rapidly, and Data Scientist is currently a position much in demand in the job market. All this creates vast research opportunities, as well as the necessity to better understand how to prepare people as researchers and professionals having the background and skills to keep active in a difficult to anticipate future.

While there are courses on Data and Information Visualization described in the literature, as well as recommendations by the SIGGRAPH Education Committee, they do not concern Data Science Programs and thus may not be entirely adequate to this type of Program. Besides the general concepts and methods usually addressed, a Visualization course tailored for this particular audience should probably emphasize specific techniques, tools, and examples of using Visualization in several phases along the Data Science process; moreover, it is reasonable to expect that new approaches, useful in practice, will be proposed by the Visualization research community that should be addressed in such a course. Likewise, the bibliography and teaching methods could probably be adapted.

We have analyzed over forty MSc Data Science programs offered in English worldwide, and the Visualization courses most of them include, and we argue that there is a need to adapt existing recommendations and create guidelines for these courses. This panel intends to debate this topic and identify issues that need further reflection.

.

Categories and Subject Descriptors (according to ACM CCS): Human-Centered Computing [Computer and Information Science Education]: Information Visualization—

1. Introduction

Data Science has been widely used to support activities in diverse domains as Science, Health, Business, and Sports, to name a few [KT18]. The underpinning theory and practice have been evolving rapidly, and Data Scientist is currently a position much in demand in the job market. All this creates vast research opportunities, as well as the necessity to better understand how to prepare people to become researchers and professionals having the background and skills to keep working in a difficult to anticipate future.

On one hand, using Visualization in several phases along the Data Science process may be most beneficial [KT18] [VW18] and has been studied for more than a decade (e.g. [TM05], [VP08], [CE13] [ZF14] [GKBP15] [CPM*15]. It is an active research topic, expected to continue so for the next years, for instance to assist in using deep learning methods, which have seen swiftly expanding application in recent years [HPC18]. On the other hand, Visualization education for Data Science practitioners is still an open topic. While there are already many courses on Data and Information Visualization, as well as books that may support them (e.g. [Maz09] [WGK10] [War12] [Mun14] [Spe14]), a significant number is offered in the scope of Computer Science, Engineering or Information Programs and thus may not be entirely adequate to Data Science Programs. Besides the general concepts and methods usually addressed in a course on Visualization, a course tailored for this particular audience should probably emphasize specific techniques, tools, and examples. Also the bibliography and teaching methods should probably be adapted. In the last decade many Universities have started to offer Data Science Programs as a response to the rapid increase of demand concerning Data Science positions (including online programs offered by renowned Universities) [O'N14]. In 2016 Tang and Sae-Lim [TSL16] conducted an exploratory content analysis of 30 Data Science programs offered at the United States from several disciplines (as Computer Science, Engineering, Mathematics and Statistics, iScholls, Business), based on the information available online. They examined the coverage of skills frequently mentioned in program descriptions and data scientist jobs: communication, information, mathematics and statistics, and visualization, and found significant gaps in current Data Science, also concerning visualization. We looked at Data Science programs offered worldwide at Universities, as well as programs in non-traditional formats, as MOOCs and professional/technical programs, and analyzed their Visualization courses in order to better

DOI: 10.2312/eged.20201033

understand what is currently is taught in those courses; however, we will focus on traditional University courses.

The rest of this proposal is organized as follows: first we present an overview of what Visualization skills seem to be necessary to a Data Scientist in practice (in industry or as a researcher); then an overview of different courses described in literature, and recommendations on a basic curriculum issued by scientific societies are summarized in sections 2 and 3; in section 4 we summarize the characteristics of 30 Visualization courses offered in the scope of Data Science MSc programs offered in English worldwide; in sections 5 specificities of Visualization for Data Science and the need for specific curricula are addressed. Finally, in section 6, the motivation of a panel on this topic is explained.

2. Visualization background for Data Science

Looking for evidence concerning the relevance of Visualization in a Data Scientist profile, as well as for information about Visualization courses in Data Science programs, we found ACM has a Data Science Task force, established by the ACM Education Council, that has been working on the role of computing discipline-specific contributions to this emerging field, and organized an academic survey and an industry survey. These surveys included questions meant to assess the relative importance of computing areas to companies, as well as in Data Science programs [DLCS19b] [DLCS19a]. The Visualization Society also surveyed data scientists in 2018 [Mee18] about issues relevant in professional data visualization. In 147 answers to the ACM Industry survey concerning the question: for the provided job title you provide to what extent do you require experience in the following areas?, about half mentioned Data Visualization as a requirement and half as elective, values comparable to those for Big Data and for Data Management Principles and Techniques. Also, in 172 answers to the academic survey, Data Visualization is offered in 25% of the cases and required in 60%. This is similar to Data Structures and Algorithms and slightly more offered and less required than Data Management Principles and Techniques. The Visualization Society survey gathered answers by 628 data scientists to questions relevant in professional data visualization, such as demographics, the most used methods and tools. Analysing these data we found that more than half of participants are relatively young (less than 35 years old), also more than half have at least a MSc degree, the great majority of the surveyed people were self-taught about Visualization, most would like to do more Visualization in future and know more about the design part of Visualization (as compared to the data part). D3 is the most used tool by the surveyed people, followed by Excel, Tableau and Python. While these surveys used convenience samples, may contain bias, and some of the questions of the ACM academic survey were concerned with undergraduate programs, we deem their results show the need of Visualization as a data scientist competence, and consequently its relevance as a subject in Data Science programs.

Searching the literature for recent studies on how analysts do their work, we found a study by Alspaugh et al. [AZLH19] who conducted interviews with thirty experienced professionals in the field of Data Science to understand how software tools support those professionals in typical exploration scenarios. While this work has limitations concerning the sample of participants and is focused on what software tools and features are considered important by these professionals, rather than on the background a Visualization course should provide to help them succeed in their work, the results suggest that contact with programming languages (as R and Python), as well as direct manipulation tools (as Tableau) may be important. Another result of this study with potential impact on a Visualization course is the recommendation that tools should record provenance and history of both data and analysis, and avoid making the user feel lack of control or visibility when automating tasks. This recommendation, confirmed by Madanagopal et al. [MRB19] in their work also based on interviews with fourteen data analysts from different domains at various expertise levels, suggests these currently active research areas may be relevant topics to include in a Visualization course.

Besides these works, we could not find any more clues based specifically on the opinion of professionals concerning what may be relevant in a Visualization course for Data Scientists. Nevertheless, the ACM SIGGRAPH Visualization Education committee [Dom15] identified a set of skill levels required for different jobs, allowing distinguish between various requirements of the job market: Visualization researcher, application oriented researcher, professional developer and professional user. We argue that these skill levels are applicable also to Data Science professionals and a bespoke Visualization course for this audience should take into consideration the needs of different positions. According to the committee, while visualization researchers perform leading edge research in core topics of visualization and their typical job markets are Universities or research labs, application oriented researchers create effective visualizations for specific application areas; their job markets are research labs and software development departments of large companies. Professional developers develop software for representations in constraint environments, and their typical job markets are software development departments and companies. Finally, professional users use visualization or graphics software, they may be able to select, to obtain visualizations selecting techniques and parameters (e.g. representations and color tables). The skills needed to succeed in the work-place in these positions may be different; however, a Visualization course should provide foundations allowing the students built the ones they need when they need them. This work is not specific to Data Scientists.

The results of the surveys, the work by Alspaugh et al. and the work about Visualization skill levels provide important clues on a Visualization course meant to prepare professionals and researchers to work in Data Science. This should be combined with previous work on Visualization courses that will be summarized in the next section.

3. Previous work on Visualization courses

We searched for work concerning Visualization education (e.g. curricula, teaching methods and cases) published in the last twenty years and found the topic has been debated in panels [GBL*05] [DKK*10] [HAK*15] and Workshops [AEHK16] and that the ACM SIGGRAPH Subcommittee on Education for Visualization maintains a site with recommendations [Dom15]; we also found several papers addressing the topic from different points of view:

debating and establishing the need of a formal education in Visualization as well as who should take such a course, and proposing curricula for Data Visualization, Information Visualization or Visual Analytics courses [Dom00] [RDD12] [San00] [RWE04] [EE12] [Ker13]; integrating realistic, but controlled problems, in Visualization courses [WNE*09]; addressing the evolution of Visualization courses [ODE*13] and describing specific methods and issues used in teaching such courses [Dom09] [SAS11] [BSO*16] [SFD16] [HA17]. Most addressed issues are relatively independent of the specific type of Program in which Visualization is taught and may be applied to the Visualization courses offered in the scope of Data Science Programs, however, we noticed an evolution in the audience of these courses. In the 80s they were typically offered to computer science and engineering students addressing scientific visualization topics as volume, flow, and terrain visualization, in the 90s and early 2000s Visualization courses were more often focused on Information Visualization, but continued to be aimed at computer science students [RDD12]; more recently Visualization has become more widely used and there is a need to debate what Visualization education should be to other fields and professions. One such case is Data Science which is different from Computer Science and thus implies a different preparation, also in Visualization. Moreover, as there may be different flavors of Data Science Programs, it might also be necessary to have different Visualization courses (e.g. specific for the Social Sciences, Business, or Biology).

4. Visualization courses offered in Data Science Programs

Teaching Visualization has two main challenges: while the subject matter of a typical course is very wide and includes much more material than a standard graduate course can easily cover, as fundamentals are based on several computer science areas as computer graphics, mathematics, or human-computer interaction, it is also necessary to address principles of perception and cognition since these are vital to develop effective and useful visualization solutions [Ker13]. This will apply to all types of Visualization courses, offered in the scope of different Programs, and thus we expected to find these characteristics in the Visualization courses of current Data Science Programs. We analyzed the curricula of 47 MSc Programs in the field of Data Science lectured in English, offered at Universities worldwide (23 in Europe, 20 in North America, 3 in Australia, and 1 in Asia). One of these Programs is offered by a consortium of European Universities; the others are offered by Universities listed in the first 200 positions of the Times Higher Education Universities Rank (the great majority in the first 100 positions). Most of these programs are coordinated by Computer Science/Engineering or Informatics Departments and involve Statistics Departments. Although some have one year (11) or less (e.g. 10 months), most of these Programs have two years (23) or one and a half years (6), including several courses and a dissertation or capstone project. While not all include a Visualization course, 30 Programs offer Visualization courses, as core or as elective, and a few others include Visualization as an area to address in a practicum or colloquium. It is also noteworthy that eight Programs offer other Human-centered computing courses (as User Interfaces Design, Human Computation and Analytics and Human-Computer Interaction). Analyzing the 30 Visualization courses, we found that the more common name is Data Visualization (13), followed by Information Visualization (6), Visualization (5) and Visual Analytics (5); we found also combinations with Exploratory Data Analysis, Data Analytics and Data Management, as well as some more specific titles (as Spatial Visualization or Large-Scale Visual Analytics).

Based on the descriptions of these courses, most seem to have similar objectives and address the general topics of a typical Visualization course; only one course includes topics more specific to Scientific Visualization (e.g. Volume and Flow Visualization). We could not find any course addressing issues concerned with Visualization for Machine Learning; perhaps this is a topic that some may include in Challenges and Opportunities, explicitly listed in their descriptions. Many courses seem to include reading papers; while we were not able to access a list of recommended bibliography for most courses, the books by Tamara Munzner [Mun14], Robert Spence [Spe07], Eduard Tufte [Tuf90] [Tuf01], and Daniel Keim et al. [KKEM10] are mentioned as references for more theoretical topics, as well as the more applied books by Andy Kirk [Kir12], Scott Murray [Mun17], Antony Unwin [Spe15] are recommended to help with the process of designing and implementing visualizations. It is worth mentioning that in recent years many Visualization books have been published [RL19], which probably will soon be reflected in Visualization courses.

5. A course on Visualization for Data Science at post-graduation level

Chaomei Chen in 2005 [Che05] considered education and training as one of the "10 Unsolved Visualization Problem"; fifteen years later the Visualization community should be in a better position concerning the challenge of researchers and practitioners learn and share principles and skills of visual communication and semiotics. Concerning another challenge enunciated by Chen related to the need of programs be linked to Universities advanced development and research efforts to consolidate the field s theory, we agree that, while not supposed to prepare for independent research, MSc programs should provide an introduction to research, and thus courses should provide a level of education allowing students to start conducting research in the field (even if not yet self-responsible).

In order to stay valid for long a time a course should address concepts and mathematical foundations upon which technical aspects can be built; yet, Visualization is not just mastering a set of concepts but acquiring skills. In this vein Rushmeier et al. [RDD12] argue that "to some extent, learning to author visualizations is like learning to write. Students should learn to develop and revise visualizations the way writing a section of text is revised".

A "general" course on Visualization for Data Scientist should first comply with the minimum requirements of a Visualization course. The ACM SIGGRAPH Subcommittee on Education for Visualization [Dom15] organizes the core topics eight themes containing facts about the most important aspects of visualization that should be all addressed in a course: Introduction to Visualization; Data; User and Tasks; Mapping; Representations; Interaction Issues; Concepts of the Visualization Process; Systems and Tools. The level of detail in which to present material for each theme is left for the educator to decide. Besides these fundamental topics, some skills, not derived from knowledge of core topics, are considered as necessary to develop expertise in visualization (e.g. as vector and matrix algebra and software/ hardware concepts). As mentioned the committee also identified a set of skill levels required for different jobs, allowing distinguish between various requirements of the job market: Visualization researcher, application oriented researcher, professional developer and professional user. The matrix relating skills and topics proposed by the same committee seems adequate to any Visualization course, including a course oriented to Data Science; however, it probably should also include topics more related to the application of Visualization in Data Science, as well as others that have been considered relevant more recently, for instance design thinking (the conception and planning of the artificial), as argued by He & Adar [HA17], due to the "wickedness" of the problems in Visualization (i.e. problems that may be not clear until the creation of a solution).

6. Why this panel at EG2020?

To the best of our knowledge no such panel has previously been organized at EG and the topic seems opportune as Data Scientist is a most searched for position by the industry. Moreover, research on using Visualization for Machine Learning has been very active for several years [KRS*12], and probably will continue to produce results that might have interesting industry applications, suggesting a need to update the curricula of Visualization courses.

The panelists will present their positions addressing each question posed in the introduction.The introductory remarks will be made by the authors of this proposal; each panelist will give a presentation (5-10 minutes); all panelists will be able to give a summary view at the end of the panel (2 minutes each) and the audience feedback will be solicited after the position statements and discussion will be encouraged.

Acknowledgements

This work was supported by IEETA - Institute of Electronics and Informatics Engineering of Aveiro, funded by National Funds through FCT, in the context of the project [UID/CEC/00127/2019].

References

[AEHK16] Adar E., Engle S., Hearst M. Joshi A., Keefe D.: Workshop on innovations in the pedagogy of data visualization. In *IEEE Visualization Conference* (2016). 2

[AZLH19] Alspaugh S., Zokaei N., Liu A.and Jin C., Hearst M. A.: Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 22–31. 2

[BSO*16] Beyer J., Strobelt H., Oppermann M., Deslauriers L., Pfister H.: Vast contest dataset use in education. In *Pedagogy of Data Visualization, Workshop at IEEE VIS* (2016). 3

[CE13] Cortez P., Embrechts M. J.: Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences 225* (2013), 1–17. 1

[Che05] Chen C.: Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications 25*, 4 (July 2005), 12–16. doi:10.1109/38.851744. 3

[CPM*15] Cashman D., Patterson G., Mosca A., Watts N., Robinson S., & Chang R.: Rnnbow: Visualizing learning via backpropagation gradients in rnns. *IEEE Computer Graphics and Applications 38*, 6 (2015), 39–50. 1

[DKK*10] Dykes J., Keefe D., Kindlmann G., Munzner T., Joshi A.: Vast contest dataset use in education. In *WisWeek2010* (2010). 2

[DLCS19a] Danyluk A., Leidig P., Cassel L., Servin C.: Acm task force on data science education. In *SIGCSE '19* (2019). 2

[DLCS19b] Danyluk A., Leidig P., Cassel L., Servin C.: *Computing Competencies for Undergraduate Data Science Curricula Initial Draft*. ACM, 2019. 2

[Dom00] Domik G.: Do we need formal education in visualization? *IEEE Computer Graphics and Applications 20*, 4 (July 2000), 16–19. doi:10.1109/38.851744. 2

[Dom09] Domik G.: Who is on my team: Building strong teams in interdisciplinary visualization courses. In *ACM SIGGRAPH ASIA 2009 Educators Program on - SIGGRAPH ASIA '09* (2009). 3

[Dom15] Domik G.: Acm siggraph curriculum for visualization (editor: G. domik, prepared by the acm siggraph education subcommittee on education for visualization). http://www.uni-paderborn.de/cs/vis, 2015. 2, 3

[EE12] Elmqvist N., Ebert D. S.: Leveraging multidisciplinarity in a visual analytics graduate course. *IEEE Computer Graphics and Applications 32*, 3 (May 2012), 84–87. 2

[GBL*05] Genetti J., Bailey M., Laidlaw D., Moorhead R., Whitaker R.: Panel 4: What should we teach in a scientific visualization class? In *IEEE Visualization Conference* (2005). 2

[GKBP15] Goldstein A., Kapelner A., Bleich J., Pitkin E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics 24*, 1 (2015), 44–65. 1

[HA17] He S., Adar E.: Vizit cards: A card-based toolkit for infovis design education. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 561–570. 3, 4

[HAK*15] Hearst M., Adar E., Kosara R., Munzner T., Schwabish J., Shneiderman .: Vis, the next generation: Teaching across the researcher-practitioner gap (panel). In *IEEE Visualization 2015* (2015). 2

[HPC18] Hohman F. M.and Kahng M., Pienta R., , Chau D. H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Computer Graphics and Applications 25*, 1 (2018), 1–20. 1

[Ker13] Kerren K.: Information visualization courses for students with a computer science background. *IEEE Computer Graphics and Applications 33*, 2 (March 2013), 12–15. 3

[Kir12] Kirk A.: *Data visualization : a successful design process*. Packt Publishing, 2012. 3

[KKEM10] Keim D., Kohlhammer J., Ellis G., Mansmann F.: *Solving problems with Visual Analytics*. Eurographics, 2010. 3

[KRS*12] Keim D., Rossi F., Seidl T., Verleysen M., Wrobel S.: *Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081 Report)*. 2012. 4

[KT18] Kelleher J., Tierney B.: *Data Science*. MIT Press, 2018. 1

[Maz09] Mazza R.: *Introduction to Information Visualization*. Springer, 2009. 1

[Mee18] Meeks E.: 2018 data visualization survey results. https : / / medium . com / nightingale / 2018-data-visualization-survey-results-26a90856476b?, 2018. 2

[MRB19] Madanagopal K., Ragan E., Benjamin P.: Analytic provenance in practice : The role of provenance in real-world visualization and data analysis environments. *IEEE Computer Graphics & Applications 39*, 6 (2019), 30–45. 2

[Mun14] MUNZNER T.: *Visualization Analysis and Design*. A K Peters, CRC Press, 2014. 1, 3

[Mun17] MUNZNER T.: *Interactive Data Visualization for the Web: An Introduction to Designing with D3*, 2 ed. O'Reilly Media, 2017. 3

[ODE*13] OWEN G. S., DOMIK G., EBERT D. S., KOHLHAMMER J., RUSHMEIER H., SOUSA SANTOS B., WEISKOPF D.: Leveraging multidisciplinarity in a visual analytics graduate course. *IEEE Computer Graphics and Applications 33*, 4 (July 2013), 14–19. 3

[O'N14] O'NEIL M.: As data proliferate, so do data-related graduate programs. *The Chronicle of Higher Education 60* (2014). 1

[RDD12] RUSHMEIER H., DYKES J., DILL J.: Revisiting the need for formal education in visualization. *IEEE Computer Graphics and Applications 27*, 6 (November 2012), 12–16. 2, 3

[RL19] REES D., LARAMEE R.: A survey of information visualization books. *Computer Graphics Forum* (2019). 3

[RWE04] ROTARD M., WEISKOPF D., ERTL T.: Curriculum for a course on scientific visualization. In *Eurographics-ACM SIGGRAPH Workshop on Computer Graphics Education* (2004), The Eurographics Association. 2

[San00] SANTOS B. S.: An introductory on visualization. *Computers & Graphics 24*, 1 (2000), 163–169. 2

[SAS11] SILVA C. T., ANDERSON E., SANTOS E.AND FREIRE J.: Using vistrails and provenance for teaching scientific visualization. *Computer Graphics Forum 30*, 1 (2011), 75–84. 3

[SFD16] SANTOS B. S., FERREIRA B. Q., DIAS P.: Using heuristic evaluation to foster visualization analysis and design skills. *IEEE Computers Graphics and Applications 36*, 1 (2016), 6–10. 3

[Spe07] SPENCE R.: *Information Visualization: Design for Interaction*, 2 ed. Pearson, 2007. 3

[Spe14] SPENCE R.: *Information Visualization: an Introduction*, 3 ed. Springer, 2014. 1

[Spe15] SPENCE R.: *Graphical Data Analysis with R*. CRC Press, 2015. 3

[TM05] TZENG F., MA K.-L.: Opening the black box - data driven visualization of neural network. In *IEEE Visualization* (2005). 1

[TSL16] TANG R., SAE-LIM W.: Data science programs in u.s. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information 32*, 3 (2016), 269–290. 1

[Tuf90] TUFTE E.: *Envisioning Information*. Graphics Press, 1990. 3

[Tuf01] TUFTE E.: *The Visual Display of Quantitative Information*. Graphics Press, 2001. 3

[VP08] VIKTOR H., PAQUET E.: Visualization techniques for data mining. In *Encyclopedia of Data Warehousing and Mining* (2008), pp. 1190–1195. 1

[VW18] VIEGAS F., WATTENBERG M.: Visualization for machine learning. In *Thirty-second Conference on Neural Information Processing Systems, NeurIPS2018* (2018), https://nips.cc/Conferences/2018/Schedule?showEvent=10986. 1

[War12] WARE C.: *Information Visualization: Perception for Design*, 3 ed. Morgan Kaufmann, 2012. 1

[WGK10] WARD M., GRINSTEIN G., KEIM D.: *Interactive Data Visualization: Foundations, Techniques, and Applications*, 2 ed. AK Peters, CRC Press, 2010. 1

[WNE*09] WHITING M. A., NORTH C., ENDERT A., SCHOLTZ J., HAACK J., VARLEY C., THOMAS J.: Vast contest dataset use in education. In *VAST 09 - IEEE Symposium on Visual Analytics Science and Technology* (2009). 3

[ZF14] ZEILER M. D., FERGUS R.: Visualizing and understanding convolutional networks. In *European Conference on Computer Vision, ECCV2014, 8689 LNCS(PART 1)* (2014). 1