Towards Open-World Scenarios: Teaching the Social Side of Data Science

Joseph Corneli[1] and **Dave Murray-Rust**[1] and **Benjamin Bach**[1]

Abstract. This article reflects on current challenges we encounter in teaching data science to graduate students. A common critique of data science classes is that examples are static and student group work is embedded in an 'artificial' and 'academic' context. We look at how we can make teaching data science classes more relevant to real-world problems. Student engagement with real problems—and not just 'real-world data sets'—has the potential to stimulate learning, exchange, and serendipity on all sides, and on different levels: noticing unexpected things in the data, developing surprising skills, finding new ways to communicate, and, lastly, in the development of new strategies for teaching, learning and practice.

1 Introduction

At first sight, data science is a hands-on technical activity, concerned with 'hard' knowledge such as statistics, machine learning, visualization, etc. But practicing data science requires an array of 'softer' skills, including understanding of the context and implications of data, communication, or collaboration. This array of requirements is reflected in common texts and references, which attempt to introduce students to the complex world of professional practice [33]; which highlight the *"need for this material to be offered more broadly"* (not just to engineering and science students) [4]; and which contrast data science teaching with *"traditional statistics courses* [...] *focused on describing techniques and their mathematical properties rather than solving real-world problems or answering questions with data"* [20].

Faced with the challenge to deliver a course on data science to graduate students in a design-oriented Master's program, we wanted to account for both 'hard' and 'soft' skills. Students came from a range of backgrounds; some with little or no prior programming experience, others with an undergraduate degree in computer science. Additionally, about 80% of the students had recently arrived from non-English-speaking countries. They brought along different cultural expectations related to communication, collaboration, and pitching.

As in other courses we've encountered, our syllabus progressed from rather closed tasks to more open ones. The first few lessons covered tutorial material on programming with Python.[2] In a second stage, we taught more applied data science problems on a specifically curated toy dataset. In a set of pre-existing csv files (detailing Titanic survivors, tips spending, etc.[3]) we systematically introduced errors such as incorrect data formatting, empty cells, spelling-errors, and non-integer values. Finding and treating these errors, as well as answering several analytical questions about the given data set was part of a second assignment. Eventually, we would connect students with larger set of data sources (Wikidata, open data from the BBC, historical databases, Twitter data, sensor data, smartphone app usage data) to develop their own approaches to analysis and visualization with less supervision.

Working together on this course lead us to discuss many challenges with current course formats and to think about methods to improve teaching the social factors involved in data science. In this short paper we reflect on our experience teaching with the above model and how we can in future improve the teaching strategy and the student experience, by including more room for serendipity in the course. We are interested in how students can encounter and cope with uncertainty, interact with people from different disciplines, and find joy in developing their skills and in noticing how these skills can shape the world around them. *How can serendipity play a role in teaching data science? How can we foster and combine engagement, discovery, and learning? How can we teach data science as a social, iterative, and mindful engagement?* The concept of serendipity can be a narrative for this kind of open-world experience: we give up some control, and this creates a real risk of failure. For example, one way to introduce serendipity into the classroom is to involve students in real-world collaborations, but this poses considerable challenges.

After enumerating and reflecting on some of these challenges (Section 2), Section 3 then surveys literature on alternative learning approaches and Section 4 talks about the role of serendipity in professional practice, comparing that with the student experience. Finally, in Section 5 we put forward our conclusions, and sum up some of the ways this work may evolve in the future.

2 Challenges in Teaching Data Science

Many current teaching setups for data science can be classified as *closed-world*, *guided*, and relatively *controlled*. These characteristics make teaching and assessment relatively straightforward, but they give an impression of data science as simply being an area of expertise, rather than a professional practice.

This section reviews the challenges we considered while developing and teaching a new course, Data Science for Design (DS4D). The following list reflects our discussions as co-developers of the course, along with our previous experience teaching data science and visualization classes, and facilitating peer and online learning experiences [10, 11, 12], as well as extended discussions with colleagues about their teaching experiences. The list is not complete: it may serve to stimulate feedback and discussion from other scholars and teachers.

C1: **Toy datasets:** the term 'toy datasets' is denotes the opposite of real-world datasets, lacking significant characteristics from the latter, e.g., size, complexity, messiness, relevance, context, etc. Toy datasets are usually small, curated, clean, and contain ground truth students

[1] University of Edinburgh, UK, contact: joseph.corneli@ed.ac.uk

[2] `http://swcarpentry.github.io/python-novice-inflammation`

[3] `https://github.com/mwaskom/seaborn-data`

are required to find. While they make assignments and assessment straightforward, they (i) require some effort from the side of the teacher (retrieval, curation, etc...), (ii) might be of little personal interest to the students, (iii) might match with available solutions from other and past courses, and (iv) might allow students to cheat by passing around their solutions.

C2: **Real-world datasets:** one way to overcome issues with toy datasets is to provide real-world data to students. Yet, real-world datasets come with their own set of challenges: (i) some may be difficult to obtain, (ii) some may be too messy to be used in a course, (iii) students might fail to comprehend the data at all, or (iv) might lack the respective knowledge to drive an analysis and interpret their findings, (v) many steps are required before analysis can take place, e.g. obtain data, transform, clean, etc.; and lastly, (vi) real-world data puts strains on evaluation and balancing difficulty.

C3: **Motivation:** Both C1 and C2 have ramifications for student motivation. Toy datasets might be too simple or just not interesting; real-world data might be too specific and not relevant to students. Allowing students to choose datasets themselves partly solves the problem but requires more preparation from the side of the teacher in terms of access, provision, description, and evaluation. However, motivation is key in learning and it exhibits multiple facets that may offset the difficulties: interests, skills, social setting, personal relevance, ideas for approaches, etc.

C4: **Complexity:** If different students use different real-world data sets, then they are likely to have widely different experiences in the course. *How do we adapt problem complexity to manageable levels?* Can a course help students learn to cope with complexity and uncertainty, phenomena they will encounter in the real world?

C5: **Relevance:** *How do students know to whom and which real-world problem their skills will be relevant?* This aspect reflects a common critique of university teaching and academia more broadly. Who is the "client"?

C6: **Soft-skills:** Since data science practitioners are not simply engaged with technical work, students need an opportunity to develop and practice relevant soft skills: problem definition, collaboration, collaboration, placing their contribution in context, understanding when and how data science can be applied, communicating their findings and discussing technical decisions with stakeholders, etc.

C7: **Method evaluation:** Eventually, every course must assess students learning outcomes. While data science is a wide field, learning outcomes will differ across courses, levels, and course audiences. *What are the learning outcomes of a course and their priority? How to evaluate each of them?* While it might be easy to evaluate technical 'hard' skills (relatively, depending on the choice of data and the methods taught), due to their nature, 'soft' skills are somewhat harder to evaluate. It would be an over-simplification to assume every student must exhibit all skills equally well.

C8: **Interdisciplinary audience:** Though not a problem in every data science course, our course was offered to related disciplines within the university and hence attracted people without programming experience and strong mathematical backgrounds. We believe interdisciplinarity in a course benefits students with technical skills and students with background in other disciplines. We believe data science is a broad methodology and serves a wider knowledge of "dealing with data".

It is probably impossible to address all of these challenges fully in a single course. Any good curriculum will balance different types of courses and learning opportunities: lectures, tutorials, projects, dissertations, presentations, etc. This gives rise to two focal questions:

- *Which structures can be implemented in individual (data science) courses in order to help weave together a consistent set of projects, skills, and engagement across courses within a curriculum?*
- *How to provide relevance and motivation in usually closed-world teaching in the context of open-world challenges?*

3 Open-World Teaching

Open-world courses, contrary to closed-world courses, are more like real-world scenarios; they can be characterized by the explicit interaction with *course-external* entities (data, collaborators, domains, etc.), less guidance, and a grain of uncertainty.

There are many ways to involve students in real-world contexts that may help address some of the above mentioned challenges. This section gives an overview of the variety of approaches that might inspire an adaption to data science classes.

Universities and Society—Various formulations of the relationship between institutions of higher learning and the wider community have been proposed and pursued. E.g., according to "the Wisconsin Idea", originated in 1905 at Wisconsin's large public "land grant" university,[4] the university must "*assume leadership in the application of knowledge for the direct improvement of the life of the people in every sphere*" [14, p. 88]. Research that adds to the store of knowledge is another fundamental obligation (*ibid.*, p. 550). Harvard takes a less interventionist stance: the university does not have a formal mission statement,[5] while its undergraduate programme states that its mission is *"to educate the citizens and citizen-leaders for our society* [via] *the transformative power of a liberal arts and sciences education."*[6]

Teaching and Research—Learning by doing research is a widespread educational practice, with various schemes available, though entrance to these is often competitive. Student involvement in research may go along with a shift from "teaching" via instruction to "peer learning" [3]. For example, recently gifted high school students have coauthored mathematics papers using online collaboration tools, with some help from mentors [19]. *Problem-based learning* involves open ended problems but, typically, a structured programme of approach [32]: it has been tried in data analytics teaching [28].

Public Action—In her proposal for a "new liberal arts" [8], Elizabeth Coleman makes contemporary social challenges the core of the curriculum. Rather than being insulated from these problems for four years, students would organize their work around challenges having to do with the environment, health, energy, economics and equity, governance, and so on. *Public action* would be adopted as a key criterion of successful performance. The relationship between students and members of the wider community is foregrounded, and practice-based education is the order of the day. As part of this effort a new Center for the Advancement of Public Action was announced [7] at Bennington and subsequently built at a cost of $20 million [26].[7]

Field-work and Collaboration—But indeed since its foundation, Bennington College had emphasized "the concrete approach" and *"engaged students in projects 'involving continuous periods in the laboratory, library, or field' under the supervision of a professor"*[39, p. 263]. *"College administrators called for education that prompted students to actively engage their social and cultural worlds"* emphasizing *"social participation and cooperation"* (*ibid.*). Similar views were expressed by other mid-20th Century thinkers (e.g., [13]).

[4] https://www.wisc.edu/wisconsin-idea/

[5] https://www.harvard.edu/about-harvard/harvard-glance

[6] https://college.harvard.edu/about/mission-and-vision

[7] Coleman's late-2000s proposal echoed aspects of an earlier contentious restructuring of Bennington College under her leadership in the 1990s [17, 24], most notably in calling for increased community engagement.

Teaching to Develop Deeper Understanding— Kenneth Burke, at Bennington in the 1940s and 50s, proposed a "'synoptic' project for 'unifying the curriculum'" [39, p. 265]. What Burke names as the "question that ultimately concerns us most" is one that can be studied by a data scientist as well as by a literature scholar: "What is the nature of a symbol-using animal?" (*ibid.*, p. 266). Other authors from the same era, working from widely different disciplinary standpoints but all influenced by ideas in cybernetics were similarly concerned with the synthesis of meaning and form (e.g., Alexander [1], Korzybski [25], Simondon [34], and von Uexküll [38]). Although our work is data-focused in name, we can nevertheless be concerned with the entire Data-Information-Knowledge-Wisdom (DIKW) hierarchy [31]—and the way meaning is made and used. Indeed, the learning outcomes in DS4D—*Data*, *Programming*, *Communication*, and *Professionalism*—are well-aligned with the terms of this hierarchy. (Furthermore, all of these issues are important insofar as we are not just teaching data science, but teaching science *per se*.)

4 Serendipity in Practice

Serendipity is linked to scientific discovery [30]. Furthermore, with today's data-driven scientific methods, *"Instead of waiting for the happy accidents in the lab, you might be able to find them in the data"* [23]. Investigators make unanticipated discoveries, find unexpected correlations, notice outliers, strange trends, etc.

Thinking about the role of serendipity in data science goes back (at least) to John W. Tukey and his definition of *Exploratory Data Analysis* (EDA) [37]. The core idea of EDA is the 'grand tour', a walkthrough of the facets and dimensions of a dataset, using a sufficiently large array of charts and data visualizations; glancing over multiple charts at once in a way that both gives a general overview over the different aspects of a data set (time, space, relations, distributions, dimensions, etc.), and also allows for serendipitous discoveries—answering questions that 'one did not know one was interested in' and which one would never have been found through a purely statistical approach. Following EDA, numerous visualization interfaces have been designed with serendipity in mind (e.g., [36, 16, 15]).

As a defined area of study "data science" dates to William S. Cleveland's more recent (2001) proposal to "enlarge the major areas of technical work of the field of statistics" [6]. Among the key elements of the proposal are the importance of work in multidisciplinary teams, and new methods for model building. He suggested that "*A basic premise is that technical areas of data science should be judged by the extent to which they enable the analyst to learn from data*" (p. 21). He remarks that "*data are the heat engine for invention*" and that "*Creative researchers, faced with problems posed by data, will respond with a wealth of new ideas that often apply much more widely than the particular data sets that gave rise to the ideas*" (*ibid.*, p. 22). He also highlighted that data science teaching should "encompass more than the university setting" and convey to non-statisticians "how valuable data science is for learning about the world" (*ibid.*, p. 24).

In the context of practices related to teaching and learning—including learning on the job—the implications of serendipity go far beyond discoveries through EDA, to the development of new professional and skill-development practices.

Consider hackathons, which bring together people with different skills, ideas, and perspectives; given a challenge these (usually interdisciplinary) teams will attempt to develop solutions in a very limited time frame. The posed challenge may require team members to develop new skills, to work with new people, and to engage with new problems. A similar idea was adopted by the IEEE VIS conference, which started a series called the *VAST Challenge*[8], which provided datasets with a specific question and a quest to solve. Participants in the challenge were entering into competition on building visualization interfaces that would visualize the data and allow people to solve the quest. Without specifically trying, hackathons can serendipitously address some of the challenges mentioned in Section 2. We can notice some common themes in open world "solutions", as found in hackathons, lab work, or data science practice, and the various teaching strategies surveyed above.

Collaboration: People with different skills may be able to find suitable opportunities for skilled practice and learn from others' skills. They may need to learn skills that foster serendipitous outcomes, taking advantage of opportunities to share early insights [9]. For example, through collaboration within and beyond the group, partnerships are formed, such as meeting talents and future employers.

Topics: In contrast to toy data sets, which are deliberately kept simple and self-contained with little connection to external knowledge, open research questions allow the possibility of serendipitous discoveries through the activation of domain knowledge and interests otherwise 'hidden' in learners. Specifically, data collaborators might help students make new connections that they would not think of on their own.

Contextualisation and interpretation: Discoveries need to be interpreted and put in context [2]. For example, learners can come up with data and insights, but only external data collaborators with the appropriate domain knowledge are able to interpret and contextualise findings from the data, eventually turning them into true discoveries. Working with domain experts helps learners to find value in their findings, and to understand any serendipitous implications of those findings.

Motivation: Learners may exercise more creativity, motivation, and interest by addressing a problem that they have chosen or helped shape, rather than a problem that got handed down to them. More broadly, Taleb advises: "Work hard, not in grunt work, but in chasing [potentially high-payoff] opportunities and maximizing exposure to them" [35, p. 110].

Skills: The talent for making serendipitous discoveries can be cultivated, and consists, in part, in learning how to pay attention to details [22]. With practice, people can get better at making interesting observations. In particular, one important skill is to discover a more interesting problem than the one you were initially working on: many new inventions were conceived by people working on some unrelated project; communication with end users can be a particularly valuable source of inspiration [22, 18].

New models, methods, organisations, and theories: As Cleveland highlights "Creative researchers, faced with problems posed by data, will respond with a wealth of new ideas" [6, p. 22]. Serendipity can apply to the discovery of new ways to think about things, not just to the discovery of facts that fit a given frame of reference.

5 Discussion and Future Work

In order to realize the concepts described in Section 4, which mechanisms for emphasising the open-world approach in data science class-rooms are needed? Again, most program curricula involve a variety of learning scenarios: open project work, lectures, tutorials, and so on, many of which contain elements of open-world teaching. E.g., writing a Master's thesis typically follows some coursework and requires students to formulate research questions, give presentations,

[8] `http://www.vacommunity.org/VAST+Challenge+2017`

plan their project, etc. Our hunch is that thinking about integrating different elements into *one single* structured course might help thinking about applying this structure to one coherent *open-world program curriculum*.

Echoing the data science pioneer Cleveland, we can say that universities are driven by an invention-engine, though they also achieve the preservation and translation of cultural values. As they learn data science, students have the opportunity to *"insert* [themselves] *into that machinery"* [29]. Accordingly, as data science teachers we are inviting students into the "power-house [. . .] of knowledge construction" [21]. We think that open-world class-projects can enhance the visibility of universities, classes, and teaching programs, and potentially make them more attractive to people pursuing continuing education.

A clear limitation of this paper is that it is based on our own experiences and discussions with colleagues. We surely need to widen the discussion, to bring in more ideas about teaching; and, eventually, we hope to provide an empirical evaluation of the methods outlined here. We hope our reflections might stimulate a pro-social approach to teaching a technical topic, one that gives soft skills due attention. We see the future of data science as inextricably wrapped up with the development of *humanistic intelligence* [27], i.e., intelligent systems with humans in the loop.

Increasingly, basic discoveries can be made using smart tools, and these tools are making inroads into interpretation of their findings: "Cognitive computing technologies can be configured to make cross-domain linkages [rather than] rely on serendipity" [5]. However, as yet, autonomous intelligent systems typically cannot deliver sophisticated, contextual, interpretations.

In spite of, and indeed, *a fortiori* because of the pace of technical advances in artificial intelligence, we need to keep in mind that teaching and doing data science requires not just technical solutions but also the cultivation of human capacities. Coleman mentions capacities for civic engagement, discrimination between core and peripheral issues, collaboration and innovation [7]. Moreover, and centrally, by expecting the expected and bringing open-world problems into the classroom, we may give students the opportunity to develop their own critical sensitivities [29].

REFERENCES

[1] Christopher Alexander, *Notes on the Synthesis of Form*, Harvard University Press, 1964.

[2] Paul André, Jaime Teevan, Susan T Dumais, et al., 'Discovery is never by chance: designing for (un) serendipity', in *Proceedings of the seventh ACM conference on Creativity and cognition*, pp. 305–314. ACM, (2009).

[3] David Boud and Alison Lee, ''Peer learning' as pedagogic discourse for research education', *Studies in Higher Education*, **30**(5), 501–516, (2005).

[4] Robert J Brunner and Edward J Kim, 'Teaching data science', *Procedia Computer Science*, **80**, 1947–1956, (2016).

[5] Ying Chen, JD Elenee Argentinis, and Griff Weber, 'IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research', *Clinical Therapeutics*, **38**(4), 688 – 701, (2016).

[6] William S Cleveland, 'Data science: an action plan for expanding the technical areas of the field of statistics', *International statistical review*, **69**(1), 21–26, (2001).

[7] Elizabeth Coleman. A call to reinvent liberal arts education. TED 2009.

[8] Elizabeth Coleman, 'Regaining the Thought-Action Continuum: A New Liberal Arts', *Yearbook of the National Society for the Study of Education*, **107**(2), 131–135, (2008).

[9] Samantha Copeland, 'On serendipity in science: discovery at the intersection of chance and wisdom', *Synthese*, 1–22, (2017).

[10] J. Corneli and C.J. Danoff, 'Paragogy', in *Proceedings of the 6th Open Knowledge Conference*, eds., Sebastian Hellmann, Philipp Frischmuth, Sören Auer, and Daniel Dietrich, Berlin, Germany, (2011).

[11] J. Corneli and A. Mikroyannidis, 'Crowdsourcing education: A role-based analysis of online learning communities', in *Collaborative Learning 2.0: Open Educational Resources*, eds., Alexandra Okada, Teresa Connolly, and Peter Scott, IGI Global, (2012).

[12] Joseph Corneli, Charles Jeffrey Danoff, Charlotte Pierce, Paola Ricaurte, and Lisa Snow Macdonald, 'Patterns of Peeragogy', in *Proceedings of the 22nd Conference on Pattern Languages of Programs*, PLoP '15, pp. 29:1–29:23, USA, (2015). The Hillside Group.

[13] Norman Cousins. Education Against Helplessness. *Saturday Review*, 19 March 1960.

[14] Merle Curti and Vernon Carstensen, *The University of Wisconsin: A History, 1848–1925*, University of Wisconsin Press, 1949.

[15] Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson, 'Visgets: Coordinated visualizations for web-based information exploration and discovery', *IEEE Transactions on Visualization and Computer Graphics*, **14**(6), (2008).

[16] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete, 'Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation', *IEEE transactions on Visualization and Computer Graphics*, **14**(6), 1539–1148, (2008).

[17] Matthew W Finkin, *The Case for Tenure*, Cornell University Press, 1996.

[18] Alfonso Gambardella, Paola Giuri, and Myriam Mariani. The value of European patents: Evidence from a survey of European inventors. Final report of the PatVal EU Project, contract HPV2-CT-2001-00013, 2005.

[19] Slava Gerovitch, Julia Braverman, and Anna Mirny. Crowdmath: Massive Research Collaboration among High School and College Students. Presented at the Enabling Mathematical Cultures Workshop, University of Oxford, 5th-7th December 2017.

[20] Stephanie C. Hicks and Rafael A. Irizarry, 'A Guide to Teaching Data Science', *The American Statistician*, (2017 [accepted for publication]).

[21] Tim Ingold, 'Anthropology is not ethnography', in *Proceedings of the British Academy*, volume 154, 69–92, Oxford University Press, (2008).

[22] Pagan Kennedy. How to Cultivate the Art of Serendipity. *New York Times*, 03 January 2016.

[23] Pagan Kennedy, *Inventology: How We Dream Up Things That Change the World*, Houghton Mifflin Harcourt, 2016.

[24] Roger Kimball, 'Bennington Lost', *Linguafranca, The Review of Academic Life*, **4**(5), (July/August 1994).

[25] Alfred Korzybski, *Science and sanity: An introduction to non-Aristotelian systems and general semantics*, IGS, 1958.

[26] Patrick McArdle. Bennington College opens new $20M educational building. *Rutland Herald*, 19 July 2011.

[27] Marvin Minsky, Ray Kurzweil, and Steve Mann, 'The society of intelligent veillance', in *Technology and Society (ISTAS)*, pp. 13–17, (2013).

[28] Miguel Núñez-del Prado and Rosario Goméz, 'Learning data analytics through a problem based learning course', in *World Engineering Education Conference (EDUNINE), IEEE*, pp. 52–56. IEEE, (2017).

[29] John Protevi, 'Preparing to learn from *Difference and Repetition*', *Journal of Philosophy: A Cross-Disciplinary Inquiry*, (2010).

[30] R. M. Roberts, *Serendipity: Accidental Discoveries in Science*, June 1989.

[31] Jennifer Rowley, 'The wisdom hierarchy: representations of the DIKW hierarchy', *Journal of information science*, **33**(2), 163–180, (2007).

[32] John R Savery and Thomas M Duffy, 'Problem based learning: An instructional model and its constructivist framework', in *Constructivist Learning Environments: Case Studies in Instructional Design*, ed., B. Wilson, 31–38, Educational Technology Publications, (1996).

[33] Rachel Schutt and Cathy O'Neil, *Doing data science: Straight talk from the frontline*, O'Reilly Media, Inc., 2013.

[34] Gilbert Simondon, *L'individuation à la lumière des notions de forme et d'information*, Éditions Jérôme Millon, [1964] 2005.

[35] Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2007.

[36] Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale, 'The bohemian bookshelf: supporting serendipitous book discoveries through information visualization', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470. ACM, (2012).

[37] John W Tukey, 'Exploratory data analysis', (1977).

[38] Jakob von Uexküll, 'The theory of meaning', *Semiotica*, **42**(1), 25–79, ([1940] 1982).

[39] Scott Wible, 'Professor Burke's "Bennington Project"', *Rhetoric Society Quarterly*, **38**(3), 259–282, (2008).