Data Science Pedagogy to Support Industry, Governmental, and Research Initiatives

Kevin Dick
Hoda Khalil
Gabriel A. Wainer
kevin.dick@carleton.ca
Systems & Computer Engineering, Carleton University
Ottawa, Ontario, Canada

Abstract

Data Science practices are increasingly leveraged in disparate domains of research, whether as part of industry workflows, governmental department initiatives, or open problems within academic communities. Herein, we describe designing term-projects to introduce senior undergraduate students to applied Data Science research for industry, governmental, or academic "clients" through a series of course assignments and client meetings. We outline the lessons learned and describe how they may be adapted within similar courses. Students are familiarized with data science best practices, obtain applied research experience, and (potentially) professionally benefit from an actual research contribution in the form of a peer-reviewed conference publication; at time of writing, we have published three student-led projects in the proceedings of eminent peer-reviewed conferences. We highly recommend introducing undergraduate students to such client-serving research applications early in their program to encourage them to consider pursuing a research-focused career path.

CCS Concepts: • **Applied computing** → **Document management and text processing**; **Enterprise data management**; **Collaborative learning**; • **Computing methodologies** → *Supervised learning by classification*; • **Software and its engineering** → **Software creation and management**.

Keywords: data science pedagogy, experiential-based learning, low-resource computing, open-source, research-centric coursework

SPLASH-E '22, December 05, 2022, Auckland, New Zealand

ACM ISBN 978-1-4503-9900-5/22/12…$15.00
https://doi.org/10.1145/3563767.3568130

ACM Reference Format:
Kevin Dick, Hoda Khalil, and Gabriel A. Wainer. 2022. Data Science Pedagogy to Support Industry, Governmental, and Research Initiatives. In *Proceedings of the 2022 ACM SIGPLAN International SPLASH-E Symposium (SPLASH-E '22), December 05, 2022, Auckland, New Zealand.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3563767.3568130

1 Introduction

The practise of data science is steadily being adopted within industry organizations, governmental departments, academic disciplines, and is itself an emergent domain of scholarly research. The scarcity of data scientists to acquire, transform, analyze, and extract meaningful insights from data is the main challenge facing data-inspired analysis in various organizations [6]. The "analytics industry" is a leader in producing analysts specialized in data-driven decision support systems, however these trainees are typically over-specialized to one particular domain and therefore lacking the transferable skills that come from a foundational understanding of data science practices [2].

In general, society would be better served by training data scientists with a well-structured and generalized talent development that imparts the utilization of computational and statistical methods and fosters the ability to then apply these methods to a target application domain [4]. Moreover, early student exposure to the fundamentals of data science best practises and curiosity-driven projects enables students to gravitate towards research-centric experiences as part of their training experiences. As described in similar work [7], the tailoring of open-ended pedagogical material to inspire creativity and innovation among students is rewarding for both student and teaching leadership alike. In fact, novel solutions and/or insights into contemporary research problems may also represent contributions worthy of peer-reviewed publication in the proceedings of relevant conferences and/or journals.

The field of data science is one that leverages both quantitative and qualitative methods to extract knowledge from (typically large) data sources with the intent to solve relevant problems and/or predict/explain outcomes [13]. A data scientist is a trained programmer with both statistical and

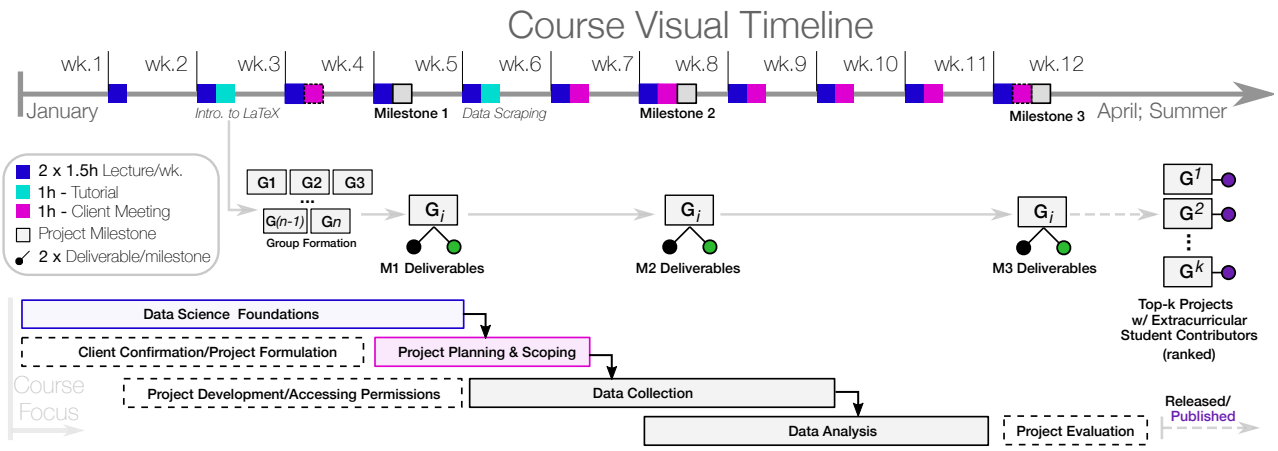

Figure 1. Conceptual overview of term-project design, milestones, & client meetings.

domain-specific expertise to effectively acquire, transform, integrate, model, and generate insights from data. Consequently, a data scientist's training must fundamentally be interdisciplinary and domain-agnostic [2].

In this article, we focus on describing the term-project design while the complete course curriculum is out of the scope of this paper. The project's assignments are designed to train senior undergraduate students, in primarily engineering programs, on the fundamentals of data science. In particular, we demonstrate that the pre-/early-course recruitment of potential "clients", project-focused pedagogy, and extracurricular refinement of course-based projects for submission to a peer-reviewed conference is highly beneficial to all project stakeholders. This term-project design is outlined in figure 1 and further described below.

2 Applied Research-Focused Didactics

In designing the requirements of the term-project, we sought to prioritize applied client-based and research-focused assignments. Our intention was to simultaneously impart a technical understanding of data science fundamentals in conjunction with an applied project putting into practise the acquired skillset. With the recent emergence of fully online data science and machine learning competition frameworks such as Kaggle, data science practitioners form digital communities to crowd-source solutions to contemporary industry, governmental, and/or research challenges [12]. This formal extra-university framework helps establish one's expertise in data science and represents a new form of credentialization that can lead to professional benefits such as employment opportunities. Thus, applied team-based projects in service to a "client" whether tailored to research or industry applications, represent excellent learning opportunities and may be leveraged for pedagogical goals.

2.1 "Introduction to Data Science"

Offered for the second time in the Winter semester of 2022, a course entitled "Introduction to Data Science" was offered as an elective course to upper-undergraduate students within the Systems and Computer Engineering department at Carleton University. The course was offered to approximately 40 enrolled students in several engineering disciplines including Software, Computer Systems, Communications, and Biomedical and Electrical engineering. The course assumed a number of prerequisites including an undergraduate level of probability and statistics and knowledge of at least one programming language. The Python programming language was used throughout the course however experience in programming was quite diverse; software engineers were clearly the most prepared at one end of the spectrum and with biomedical engineers at the other end having only an introductory programming course in their program.

Student project groups comprised 4-6 members and were self-organized according to project subject interest. Prior to the course, we performed outreach to potential industry, governmental, and academic "clients" for which an applied project collaboration could be established. Given the limited number of secured clients (only two for the latest cohort; however one client provided three potential projects), a large number of diverse research-centric projects were also provided to students; here, the teaching leadership represented the "academic client". Finally, students were permitted to define their own project in consultation with the teaching leadership. All project ideas were advertised on a course-specific GitHub repository.

In Figure 1, we depict a conceptual overview of the components of such a term project in a data science course. The 12-week course featured approximately three hours of weekly lectures covering theoretical concepts, complimented by occasional hour-long tutorials and occasional client-based meetings that were highly-variable dependent

Table 1. Overview of Project Topics and Team Composition from the 2022 Cohort

Team ID	Team Size	Client	Specialized Topics	Brief Description
G1	5	Industry Partner	ML, NLP, Knowledge Graphs	Automatically searching online for company mergers & acquisitions and building a representative knowledge graph.
G2	6	Gov. Agency	GIS, Transit Analysis, Statistics	Collection of all Public Transit feeds from 200+ cities/municipalities throughout Canada to study transit availability.
G3	5	Academic Client	Multi-year, Multi-region Census Analysis, Statistics	Identifying the "Modal Canadian" throughout Canada, across census regions, and over census years.
G4	4	Academic Client	NLP, Sentiment Analysis, ML Inference	Mapping concepts/themes of hateful tweet content from the Canadian Twittersphere.
G5	6	Academic Client	Crypto Market Analysis	Creation of a multi-perspective Crypto recommendation system integrating both market performance and NLP-based measures.
G6	4	Academic Client	Crypto Market Analysis, Reddit NLP Analysis	Creation of a Crypto (alt)coin market performance predictor based on Reddit sentiment analyses.
G7	6	Academic Client	Crypto/NFT Market Analysis, Twitter NLP Analysis	Large-scale correlation study of NFT collection market performance using Tweet-derived measurements.
G8	4	Academic Client	Amazon Product Pricing, Data Scraping	Large-scale correlation study of Amazon product price variation during natural disaster events.

on client availability and expectations (*e.g.* regular weekly meetings at one extreme and twice throughout the semester at the other). These client meetings represented opportunities for the students to present progress on their work, obtain direct guidance, and ultimately, deliver the entirety of the data, codebase, analysis, and technical report to the client. As the project topics spanned a wide array of application domains and "flavours" of data science (Table 1), students were provided with flexible template materials and milestone instructions that could be broadly applied regardless of the specific implementation of their project[1]. Additionally, the project scope and expectation were regularly refined in consultation with students and according to available resources.

2.2 Navigating the Jungle of Online Data Sources

Excitingly, data science methods can be flexibly applied to a wide variety of domains given the appropriate (systematic) collection and representation of data by abstracting domain-specific considerations of a problem into a representation amenable to a typical data science analysis pipeline. Consequently, students may develop their intuition for data acquisition, transformation, and modeling to then apply themselves to domains in which they might not otherwise have taken interest. Promisingly, this enables data science pedagogy to be domain-agnostic and adapted to a diversity of client-based problems.

Fundamental to all projects is the data quantity and quality itself. Our course places much emphasis on the ethical acquisition (*e.g.* scraping) of online data through adherence to best practices [9]. For general online data scraping projects, the instructor and the project coordinator provided flexible guidance on how best to design and execute scrapers, generally requiring the assessment of feasibility/**ethicality** in the jungle of potential online sources. For Natural Language Processing (NLP) projects, the teaching leadership applied for "academic" licences to Twitter APIs on behalf of the teams leveraging those data as part of their work. Client-based projects necessitated the negotiation of data to respect client data management policies. In all cases, this represents considerable project overhead that must be flexibly managed by the teaching leadership and is independent of the actual course instruction. Typically managed in the first month of the course (Figure 1; M1-M2), a dedicated *Project Coordinator* among the teaching leadership is required. Thereafter, the role shifts to compute and analysis support.

2.3 Google Collab Supports Large-Scale Computation

The large-scale analysis or inference of ML models on novel data typically requires high-performance computing/GPU access. To provide all students with an equal computing platform, Google Collab was recommended and demonstrated as an adaptable environment. As a free cloud-based development environment, students, in low computational resource environments are enabled to contribute to research initiatives with compute resources they might not otherwise be able to leverage.

For projects requiring GPU resources to develop and train ML models and/or generate model inferences on large-scale datasets (*e.g.* sentiment analysis on social media posts), the free-tier Google Collab allocations considered across teams of 4-6 members enabled general parallel computing opportunities. Consequently, novel datasets could be acquired,

[1] https://github.com/chazingtheinfinite/intro-to-data-science

merged, and transformed to generate meaningful insights to specific research questions.

2.4 Analyses to Contribute to Client-Specific Data Science Problems

Project-specific outcomes were generally outlined and incrementally refined through each milestone. Since all subsequent data analyses depended upon the acquired data, the breadth and depth of analysis were modulated in accordance to the resultant data.

For the G2 project in collaboration with a government agency, the manual data collection of public transit feed sources represented a considerable investment in generating novel dataset which was the client's desired deliverable. Consequentially, the data analysis itself only represented surface-level data summaries.

At the other extreme, the G1 project, in collaboration with an industry partner, impressively produced a fully generalizable ML framework for task-agnostic knowledge graph generation; succeeding both in achieving the client's intended outcomes while developing a configurable tool readily usable by research communities, now recently published [11].

Projects G5-G7, interestingly, each related to studying facets of the cryptocurrency and non-fungible token (NFT) markets, reflecting the general interests of students around the hype of crypto/alt-coins/NFTs prior to the May 2022 crash of these markets [10]. In each case, projects were focused to a specific set of research questions in an attempt to elucidate various patterns in cryptocurrency markets.

While ambitious, we expected that framing the project milestones as an effort to investigate novel research questions would foster creativity and generate interest in research applications in general. Moreover, the suggestion that the work produced within the context of the term-project might eventually represent research contributions in collaboration with their client further motivated student groups. Emphasis on reproducibility ensured that resulting datasets and analysis pipelines could be leveraged by the client and research communities at large.

3 Comparison to Related Pedagogical Initiatives

The course design we propose may share many commonalities to related initiatives in other fields. Most notably, the prototypical Master of Business Administration (MBA) program features "case studies" that may pursue similar learning outcomes [3], however, the data science thematic of our course design aims to incorporate novel and creative technocentric approaches that might not otherwise be considered in a traditional or structured program.

With a broad and flexible view to incorporate projects originating from industry, gouvernmental, and/or academic initiatives, this course design is adapted to the broad interests of the participating students. In essence, these newly defined projects represent experiential learning in tertiary education with project opportunities adapted to the interests of the ever-evolving participating class and in tune with emergent trends of research interest. Akin to these projects are final year/capstone projects for which a wealth of pedagogical research currently exists however the presented projects may not sufficiently cover a diversity of industry, governmental, and/or academic research initiative [8]. Similarly, research-oriented undergraduate projects expose students to academic research, however, they may be limited in the scope of exposure to engaged industry or gouvernmental clients.

4 Lessons Learned

Lesson 1 - Research Experience Inspires Interest. Similar to previous experience [7], some students considered pursuing research for the first time. For the majority of the students, this course is offered in the terminal semester of their program, leaving little opportunity to tailor their undergraduate experiences towards research-focused interests. Interestingly, the G1 student group of the 2022 cohort comprised solely 3rd-year students in the penultimate program year and who expressed interest in furthering their research experiences prior to graduation.

From these client-focused experiences, students realized that provided access to cloud-based resources, scraping a novel dataset, and research-focused guidance, they could contribute to interdisciplinary challenges with relative ease. Our advice to data science educators is to focus on generalized data science fundamentals and encouraging literacy across multiple application domains to ensure students can work in conjunction with various subject matter experts and stakeholders.

Lesson 2 - Organize Interesting and Well-Formulated (Client-Specific) Problems. To engage students all projects (whether client-related or not) were tailored to an open problem intended to pique the interest of students. Interestingly, the government agency offered three putative projects with only G2 being selected. Conversely, a cryptocurrency-related project was suggest by the teaching leadership and three groups (G5-G7) each developed unique project proposals within the crypto-fintech domain. Ultimately, we learned that students, when provided open-themed projects, will gravitate towards data-driven projects that are most likely to be trendy and/or benefit them personally/professionally. The project potential for peer-reviewed publication and thus professional benefit is likely too nebulous to play into their decision-making early in the course.

Lesson 3 - A Part-Time Interdisciplinary Project Coordinator is Essential. To sufficiently support multiple projects spanning numerous application domains, the teaching leadership must comprise at least one project coordinator. This may be a mature teaching assistant, project manager,

and/or research staff to coordinate project topics and represent the "academic client" for projects without industry and/or governmental collaborators. While project topics are prearranged and left to the students to select on a first-come-first-served basis, there exists a risk that students whose project comes from the project coordinator and not an actual client may feel slighted that they have a simulated experience as compared to their peers. In these instances the proposed projects are focused on yet-unaddressed research problems to stimulate engagement.

In the recent offerings of the course, an interdisciplinary mature PhD candidate assumed this role in support of a postdoctoral course instructor with significant industry experience. To ensure that student-based projects are provided sufficient extracurricular support, we recommend that a dedicated part-time interdisciplinary project coordinator help manage course-specific projects and the subsequent extracurricular peer-review publication process along with the course instructor and/or a project-sponsoring department professor.

Lesson 4 - Individual Project Scope must be Flexible. As previously discussed, a client-based and research-focused term-project must adapt to the various obstacles resulting in any of the projects. Non-exhaustively, these may be due to student-specific limitations, challenges in data collection, limited computing resources, erroneous methodologies, *etc.* Thus, great creativity and flexibility is required of the teaching leadership to successfully support students and navigate the project towards a meaningful outcome. At times, this may require direct intervention of the project coordinator and the course instructor to navigate student groups around a particular obstacle. In such a limited time-frame, students are not expected to achieve mastery in specific data science applications and thus the redefinition of project scope and/or intervention of the teaching leadership are intended to circumnavigate challenges and exemplify cohesion of a prototypical data science team supervised by senior members.

Lesson 5 - Extracurricular Research Investment is Necessary for Peer-Reviewed Publication of Findings. Despite formulating the technical reports as closely resembling the standard format and structure of conference proceedings, by the course conclusion the final project reports represent only premature versions of what would ultimately be submitted for conference consideration. We emphasize that this course be run in the Winter semester of the academic year to allow for the general flexibility of students through the subsequent summer months. Our previous experiences has indicated that, when presented with the opportunity to extend their work, a subset of the student groups are willing to volunteer their own time to improving the work whilst managing summer co-op work or full-time employment. As depicted in Figure 1, the selection of candidate projects for peer-reviewed publication requires the qualitative "ranking" of projects according to their peer-reviewable "publication potential" based on the course grade, past experiences of the teaching leadership, and in consultation with the client, if any. The final report and codebase is then critically revised to summarize necessary modifications and students are engaged to determine interest in pursing publication. Subsequent student/client engagement on the publication pipeline is then ultimately rewarding to assist all stakeholders in seeing their work realized in an official venue. To date, this approach has resulted in three peer-reviewed conference publications with shared first-authoring students [1, 5, 11].

Lesson 6 - Promote Success & Research Outcomes. An important realization of this term project design is that course didactics are mutually beneficial to student, client, and teaching leadership alike. Students gain invaluable research experience in applying their recently-acquired data science knowledge to a relevant client-specific problem. Given the emphasis on reproducibility throughout the course, the systematic collection, analysis, & dissemination of novel datasets can produce insights into contemporary problems.

A publication plan should be considered early as part of the project organization with students authors each sharing scientific credit as alphabetized co-first-authors denoted with an asterix for "equal contribution", and the course teaching leadership (and collaborating client, if any) representing the terminal supervising author(s). Given that research translation and peer-reviewed publication is beyond the scope of a 12-week undergraduate project, this effort resides purely with the teaching leadership, however, it represents a unique opportunity to promote student success and assist in advancing the careers of participating TAs & students; co-authorship on a published peer-reviewed article for undergraduate coursework is rare.

Lesson 7 - Intra-University Financial Support is Necessary. Regardless of client engagement, securing funding for the potential publication of student-generated work is challenging. Industry partners and governmental agencies (without established collaborative agreements) are typically limited in their ability to support initiatives outside of their respective organizations. Thus, to financially support this extracurricular publication of such work intra-university funding must be secured. Fortunately, our past student-centric initiatives have been supported by a generous departmental faculty sponsor. To expand such initiatives in a sustainable manner, we recommend that universities allot funding pools to encourage and support student-based data science publications.

5 Conclusion

Leveraging student creativity in generating novel datasets with notable analytical outcomes, course educators, as with their students and clients, are expected to benefit from the experience of tailoring contemporary data science projects to relevant applied (research) projects. We encourage data science educators everywhere to follow from our example.

References

[1] Rahul Anilkumar, Benjamin Melone, Michael Patsula, Christophe Tran, Christopher Wang, Kevin Dick, Hoda Khalil, and Gabriel Wainer. 2022. Canadian jobs amid a pandemic: examining the relationship between professional industry and salary to regional key performance indicators. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, IEEE, Torino, Italy, 235–240.

[2] Daniel Adomako Asamoah, Derek Doran, and Shu Schiller. 2020. Interdisciplinarity in data science pedagogy: a foundational design. *Journal of Computer Information Systems* 60, 4 (2020), 370–377.

[3] Valentina Chkoniya. 2021. Success Factors for using case method in teaching applied data science education. *European Journal of Education* 4, 1 (2021), 76–85.

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

[5] Laura Colley, Andrew Dybka, Adam Gauthier, Jacob Laboissonniere, Alexandre Mougeot, Nayeeb Mowla, Kevin Dick, Hoda Khalil, and Gabriel Wainer. 2022. Elucidation of the Relationship Between a Song's Spotify Descriptive Metrics and its Popularity on Various Platforms. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, IEEE, Torino, Italy, 241–249.

[6] Vasant Dhar. 2013. Data science and prediction. *Commun. ACM* 56, 12 (2013), 64–73.

[7] Kevin Dick, Daniel G Kyrollos, and James R Green. 2021. Machine learning pedagogy to support the research community. In *Proceedings of the 2021 ACM SIGPLAN International Symposium on SPLASH-E*. ACM SIGPLAN, Chicago, USA, 43–48.

[8] Sandra Gorka, Jacob R Miller, and Brandon J Howe. 2007. Developing realistic capstone projects in conjunction with industry. In *Proceedings of the 8th ACM SIGITE conference on Information technology education*. 27–32.

[9] Alex Luscombe, Kevin Dick, and Kevin Walby. 2022. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity* 56, 3 (2022), 1023–1044.

[10] Joel Seligman. 2022. The Rise and Fall of Cryptocurrency: The Three Paths Forward. *Washington University in St. Louis Legal Studies Research Paper* 22-06 (2022), 01.

[11] Nicholas Sendyk, Curtis Davies, Titus Priscu, Miles Sutherland, Atallah Madi, Kevin Dick, Hoda Khalil, Ala Abu Alkheir, and Grabriel Wainer. 2022. A Task-Agnostic Machine Learning Framework for Dynamic Knowledge Graphs. *CASCON '22: Proceedings of the 32nd Annual International Conference on Computer Science and Software Engineering* (2022).

[12] Christoph Tauchert, Peter Buxmann, and Jannis Lambinus. 2020. Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions. In *Proceedings of the 53rd Hawaii international conference on system sciences*. https://doi.org/10.24251/HICSS.2020.029

[13] Matthew A Waller and Stanley E Fawcett. 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. , 77–84 pages.

Received 2022-08-26; accepted 2022-09-30