Big Data
Volume 7, Number 4, 2019
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2019.0100

ORIGINAL ARTICLE

An Experience-Centered Approach to Training Effective Data Scientists

Kit T. Rodolfa,[1,2] Adolfo De Unanue,[1,3] Matt Gee,[1] and Rayid Ghani[1,2,4,*]

Abstract

Like medicine, psychology, or education, data science is fundamentally an applied discipline, with most students who receive advanced degrees in the field going on to work on practical problems. Unlike these disciplines, however, data science education remains heavily focused on theory and methods, and practical coursework typically revolves around cleaned or simplified data sets that have little analog in professional applications. We believe that the environment in which new data scientists are trained should more accurately reflect that in which they will eventually practice, and we propose here a data science master's degree program that takes inspiration from the residency model used in medicine. Students in the suggested program would spend their time working on a practical problem with an industry, government, or nonprofit partner, supplemented with coursework in data science methods and theory. We also discuss how this program can also be implemented in shorter formats to augment existing professional master's programs in different disciplines. This approach to learning by doing is designed to fill gaps in our current approach to data science education and ensure that students develop the skills they need to practice data science in a professional context and under the many constraints imposed by that context.

Keywords: data science; education; training; curriculum; data science competencies

Introduction

A century ago, whether a patient's doctor had received any practical experience during their medical schooling was a haphazard function of the whims of their instructors. According to Long[1]:

> Some professors required virtual indentured servitude, whereas the educational experiences offered by others were too brief to be of value. Indeed, young surgeons might never have performed an operation.

With the advent of the residency model, medical training not only became increasingly standardized but also ensured young doctors gained critical practical experience working with patients and taking on escalating responsibility over time. Although residency has been established in medicine for so long no data seem to exist to speak to its impact, its adoption is generally credited with revolutionizing the training and competency of new doctors. Adoption of similar models for students in psychology and, more recently, education[2,3] further illustrate the importance of hands-on training in practical fields.

Just as the field of medicine reflects the practical applications of basic research in biology, physiology, and anatomy, the field of data science is primarily concerned with the practical application of knowledge developed in the more theoretical fields of computer science, machine learning, and statistics. Unlike medicine, however, graduate education in data science remains heavily focused on these theoretical underpinnings.

For the purposes of this article, we define data science as a broad set of computational and data-driven methodologies and techniques coming from computer science, statistics, math, and social sciences, including machine learning and artificial intelligence used to

[1]*Center for Data Science and Public Policy, Computer Science Department and Harris School of Public Policy, University of Chicago, Chicago, Illinois.*
[2]*Machine Learning Department and* [4]*Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania.*
[3]*Department of Mathematics, Instituto Tecnológico Autónomo de México, Mexico City, Mexico.*

*Address correspondence to: *Rayid Ghani, Machine Learning Department and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, 4902 Forbes Avenue, Gates Hillman 8023, Pittsburgh, PA 15213,* E-mail: rayid@cmu.edu

tackle real-world problems. Our audience here is students interested in learning how to solve problems using data science and not necessarily develop new methods.

The extent and quality of students' work with practical real-world data (if any) during their education is highly dependent on their particular degree program and often cited as an important gap in the current educational environment.[4–7] While the cumulative exercises (or capstone projects) employed by some programs can provide students with a valuable introduction to the nuances and challenges of working with real data, some have voiced concerns that these are not consistently implemented,[8,9] and we contend that developing these practical skills and intuition requires sufficient time and experience that it should live at the core of data science degree programs.

As a result, many students entering industry or the public sector with a master's degree in data science may have strong foundations in understanding the technical details of a wide variety of methods and techniques, but little experience in choosing the right method for a given problem, applying it to messy real-world data, solving a problem with real-world constraints, and explaining the results of their work to decision makers and stakeholders. Deficits in communication abilities and relevant experience with practical problems were highlighted among the critical skills for new hires in an employer survey performed by an American Statistical Association working group on Master's education in statistics,[10] whereas other researchers are advocating for the inclusion of more practical data science skills in the undergraduate and graduate curriculum.[4,11–16]

As with other professional disciplines, these practical skills also need to be grounded in a strong foundation of professional ethics and an understanding of existing legal regulations around data collection and use. The history of medicine again provides some helpful context: in his seminal work on the development of medicine as a profession in America, Starr[17] described self-regulation and a code of ethics with an enshrined service orientation as defining characteristics of professional sovereignty. Regardless of whether they work in the public, private, or nonprofit sector, data science practitioners will find themselves facing the legal and ethical implications of their work on a regular basis, whether in the form of requests from supervisors to focus on certain aspects of their data, the potential for disparate impact of their predictive models, or the uses and protection of individuals' sensitive data. The recent interest and active conversation around the nature of professional ethics for data scientists[18–21] only serves to highlight how critical it is that students entering this field are well equipped with a toolkit that helps them navigate these issues. Any program seeking to train data scientists would be remiss to treat professional ethics as anything less than a central component of their curriculum.

Recently, a number of innovative degree programs have begun to seek ways to incorporate deeper practical experience relevant to employers (including master's programs at Northwestern[22] and Boston University[23]) or broader cross-discipline curriculum (for instance, in a new data science institute and undergraduate program at UC San Diego[24] and cross-department degree programs at institutions including the University of Chicago[25] and Carnegie Mellon University[26]).

Here, we extend these ideas to propose a new model for graduate education in data science, drawing inspiration from the medical residency and centered around a prolonged applied project as the focal point of the degree. A degree program focused on gaining real-world experience will be better aligned with ensuring that students develop the skills they need to be successful data science practitioners as they begin their careers. While we believe coursework highlighting the theoretical basis of data science methods should remain an important part of a student's experience, (1) it should span several disciplines including computer science, statistics, and social sciences and (2) it should serve to supplement and enhance students' applied work rather than act as the primary element of the program.

Domains of Data Science Competency

Recent work in medical[1] and psychology[27] education has begun to view professional education in these areas through the lens of core competencies, providing a valuable framework for designing degree and post-degree programs that best serve the needs of students. Table 1 presents the domains of data science competency we considered while developing our proposed program, with more detail on each domain below. While a given data scientist may focus their career on a certain class of problems or methods, this list reflects a core set of foundational skills that cut across these areas of specialization.

Problem definition

The first competency domain is the ability to define and scope a problem effectively. In our experience, we

Table 1. Core domains of data science competency

Domain	*Key skills*
Problem definition	Defining problem scope, identifying goals, and choosing appropriate methods and tools
Data preparation	Cleaning, integrating, structuring, and storing messy real-world data, dealing with missingness and imputation
Modeling and analysis	Foundational methods in statistics, machine learning, and social sciences
Evaluating results	Choosing appropriate performance metrics, measuring generalization performance, model selection, experiment design
Professional ethics	Ethical conduct in the use and presentation of data and data science methods, bias and fairness in modeling, privacy and protecting sensitive and confidential data
Communication	Building trust in data, understanding organizational needs, communicating about complex methods and results, visualization, persuading people to use what you have done

have framed scoping as involving answers to the following four questions[28]: (1) Who are the stakeholders for the analysis at hand and what are their goals for the project? (2) How will the output of the analysis be used (does it support certain actions)? (3) What data (either internal or external) are available and/or needed for the project? and (4) What tools and methods are best suited for approaching the problem and how will the results of the analysis be validated?

Data preparation

Most real-world problems involve data sets that are far more messy than anything students will have encountered in traditional academic programs. Effective data scientists need to be able to understand the data at hand as well as its context and generative process. They should be experienced with ingesting, cleaning, integrating, and storage of raw data files from multiple sources, dealing with gaps and missingness in the data, and preparing the data for further analysis or modeling.

Modeling and analysis

Traditional degree programs are well suited at building one aspect of this competency domain, which is developing the toolkit of modeling and analytical methods. We need to augment the theoretical understanding with some key skills:

First, they need the ability to apply these methods to practical problems. Professional data scientists are best equipped when they understand the applications of these methods, their theoretical underpinnings, and crucially their limitations. Successfully applying these methods also requires sufficient domain understanding, ability to communicate with experts in that area, and the experience to turn that into a set of features/predictors that need to be provided to the modeling methods.

Second, in addition to building a large number and variety of models, data scientists need the ability to select models that are likely to perform well* in the future. Since real-world data science problems can be approached with a variety of different methods, effective model selection is essential to good data science. The model selection process needs an understanding of not only what the models are but more critically, how they will be applied: What validation methodology should be used to mimic how the system will be used? Which metrics are going to be used to evaluate them? When multiple approaches are available, how should these be compared relative to one another and the best option chosen for validation trials and eventual deployment?

Evaluating results

Successful data science projects must not only provide results that perform well on existing data but also which actually continue to perform as expected when put into practice. While we have techniques to do initial model evaluation using historical data, our true goal is to select a model that is effective in the future by answering some key questions. Does the model or analysis continue to work as expected based on its performance on the data used to develop it? If there are several candidates to put into use, how will their real-world performance be compared for the purposes of choosing a final option to deploy. Validating a model or analysis in a real setting with truly novel data is a particularly important aspect of successful data science work, often combining machine learning methods with the design, execution, and analysis of a randomized experiment (or nonexperimental pilot) as well as soliciting and integrating feedback from nontechnical end users. Likewise, the responsibilities of a data scientist do not end when a model or analysis is deployed—they must also consider how a deployed system will be monitored for performance degradation, how it will effectively interact with its users, and how outcomes of the system actually impact equity and fairness over time.

*Well = everything we care about: accurate, fair, interpretable, stable, and so on.

Professional ethics

In the context of data science, ethical conduct involves both technical and nontechnical skills. Students of data science should receive training in how to approach problems through an ethical perspective and ensure results they present reflect a fair and well-supported reading of the data. They need to understand the ethics around the use of the data they have access to, and possible implications on people affected by the system they are building. Likewise, they should be prepared for difficult situations in which they encounter institutional or managerial pressure to present a less scrupulous read of the data. Training in understanding issues around transparency of the work being performed and accountability of downstream impact is critical here. Effective data science professionals should also be well versed in technical methods for measuring bias and fairness in their results (and underlying data), as well as considering the implications of trade-offs in how these terms are defined or measured. A working knowledge of policy and legal frameworks for protecting private or sensitive data as well as research involving human subjects is also critical to the effective and ethical practice of data science.

Communication

As with ethics, communicating about data science can involve technical and nontechnical skills. Professional data scientists need to be able to understand the needs of the organization in which they work and how those needs translate into data at hand (or that could be collected). Likewise, they need to be capable of building organizational trust in the use of analytics and data to inform decision-making. A critical skill is the ability to effectively communicate about complex methods and results to nontechnical audiences, often involving methods for data visualization and developing interpretable explanations from otherwise black box models.

Data Science Project Lifecycle

While the discussion above provides a structure for thinking about the competencies required for a data scientist to be effective, these domains are far from independent of one another and are used repeated over the course of a given project. Figure 1 shows the interconnectedness of the different phases of a typical data science project, whereas Figure 2 illustrates how skill domains are put into use across these phases (with dark gray boxes indicating a primary role for the skill in that phase and lighter gray boxes indicating a secondary, but still important, role).

These phases are described in more detail below, illustrated through a real-world example of reducing lead poisoning in children.

Phase I: Problem identification

Even before the outset of a data scientist's work, stakeholders or policy makers may need to make a decision about pursuing a solution to a problem. This involves evaluating whether the problem is significant, whether it is feasible to solve the problem with data science, and whether there is (or will be) commitment internally to allocate resources to addressing the problem. Data scientists have a critical role in this process, both providing a voice about what is technically feasible and why it may provide improvements over current practices, as well as an ethical duty to highlight the limitations and risks involved. Similarly, a solid understanding of laws and best practices around data privacy and sharing can be essential to helping decision makers understand how the data they have can be used and what other data they may or may not be able to collect for the project.

Example: Public health officials identify high rates of lead poisoning in children in their jurisdiction, but current practice only remediates issues in homes after a child has tested positive for elevated blood lead levels. They would like to reduce lead poisoning in children by proactively identifying children who may be at risk before poisoning occurs.

Phase II: Scoping

As a project begins to get underway, competencies in communication and problem definition will be particularly important in scoping the actual work. The data scientist needs to be able to evaluate what questions can be answered with the available data as well as work closely with the stakeholders to understand their needs and how any models and analyses will actually be put into use.[28] Ethical concerns at this stage include considering how sensitive data will be handled and protected as well as establishing criteria by which analysis will be evaluated in ways that balance efficiency, effectiveness, and equity.

Example: A scoping session is held including public health officials, clinicians, lead hazard inspections teams, and data scientists to understand the data available and how risk scores would be put into use. Because of the need to work with private health information and

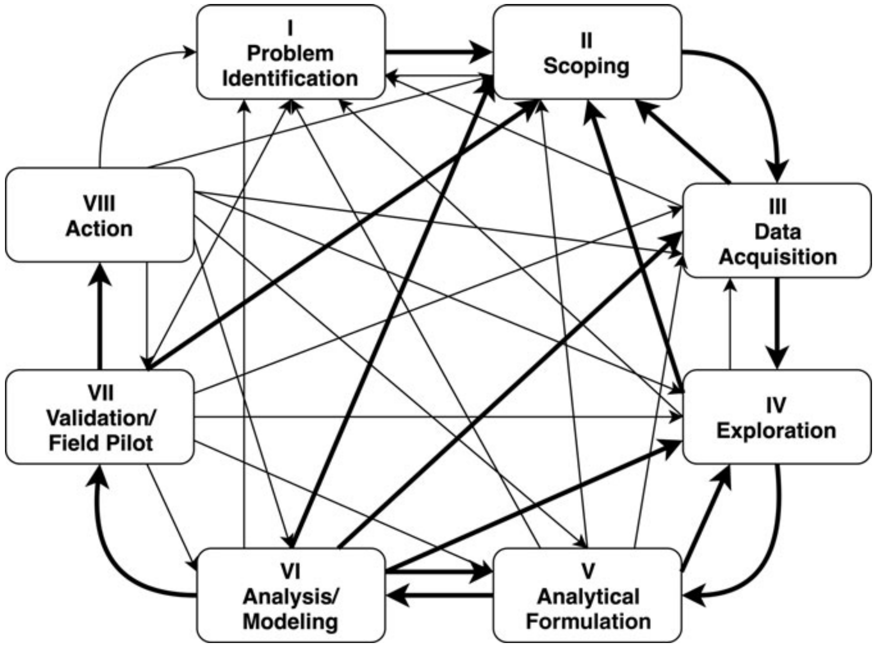
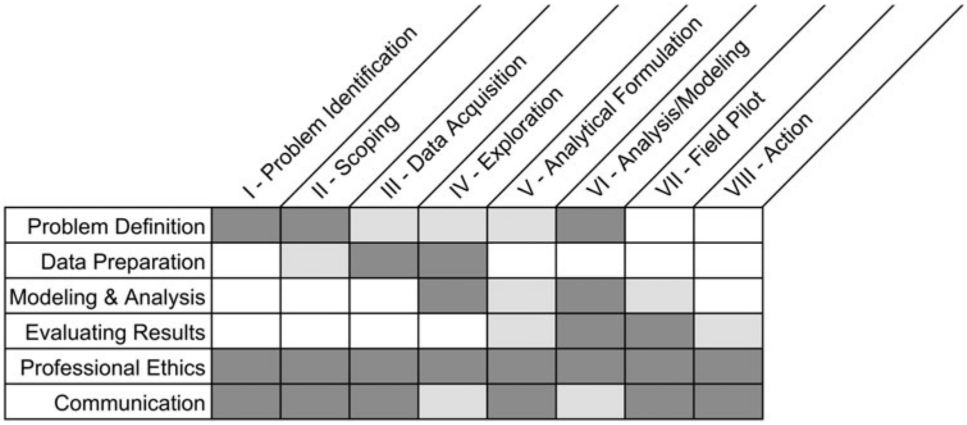


FIG. 1. Lifecycle of a data science project. The phases of real projects are highly interconnected, with insights from later phases informing and revising earlier ones, resembling more of a "web" than a typical linear pipeline or cycle. Bold arrows show more frequent transitions typically repeated multiple times in the course of a project.

data pertaining to children, the decision is made to restrict all analytical work to the Department of Public Health's secure server environment. Primary intervention is identified as lead hazard inspections in homes with high risk of lead hazards and presence of a child younger than 12 months. The key goal identified in the scoping phase was to effectively reduce childhood lead poisoning in an equitable manner across underserved communities.

Phase III: Data acquisition

Acquiring, storing, linking, understanding, and preparing data for analysis in a real-world project often entail an involved and iterative process, requiring working



	I - Problem Identification	II - Scoping	III - Data Acquisition	IV - Exploration	V - Analytical Formulation	VI - Analysis/Modeling	VII - Field Pilot	VIII - Action
Problem Definition	▓	▓	░	░	░	▓		
Data Preparation		░	▓	▓				
Modeling & Analysis				▓	░	▓	░	
Evaluating Results					░	▓	▓	░
Professional Ethics	▓	▓	▓	▓	▓	▓	▓	▓
Communication	▓	▓	▓	░	▓	░	▓	▓

FIG. 2. Competency domain use through the lifecycle of a data science project.

closely with the owners of various data sources to ensure any transferred data are provided in a consistent and reliable format and necessary steps are taken to protect private or sensitive information. During this phase of work, the data scientist needs to apply skills working with and structuring raw data to get it into a storage format that is appropriate for linking it with other data sources. Each of those steps requires active communication with the project's stakeholders to understand the context in which the data were collected and structured, its idiosyncrasies, and ensure data definitions actually describe the events they are supposed to reflect.

Example: The Department of Public Health provides a database and server for analysis in their environment with an extract of individual-level blood lead test results as well as inspection reports from lead hazard inspections. Data from additional sources are imported into the environment, including census data, childhood nutrition benefit program data (to identify potentially vulnerable children), and information about buildings from the county assessor website. Address normalization and geocoding allows data to be linked across these sources, and data scientists work closely with the owner of each data source to ensure they understand the data structures and fields.

Phase IV: Exploration

This initial phase of analysis focuses on exploring the trends and relationships in the data through summary statistics, visualization, and preliminary modeling. Although many traditional data science programs will provide students with a basic toolkit of applicable statistical methods, the open-ended nature of working with messy real-world data can be daunting to students who have only worked with highly curated data sets in a guided setting. In most projects, this stage also requires a facility with handling missing data as well as identifying potential bias and disparities in the data as well as the labels and potential features.

Example: The data scientists use a combination of descriptive statistics, bivariate correlations, spatial and temporal analysis to begin to understand the relationships in the data and its limitations. Missing values in the childhood nutrition benefit data set identify an error in the extract, transform, load process that is corrected with a new data extract, while a sharp decrease in the number of blood tests in data older than 17 years reflects a change in policy around testing that defines the limitation in historical training data.

Phase V: Analytical formulation

This phase involves formulating the initial problem as a concrete analytical problem. In most cases, a greater understanding of the available data and its nuances will result in a greater understanding of the problem itself as well. During this phase, a data scientist will need to be able to effectively communicate preliminary results to stakeholders, including any limitations or shortcomings. At the end of this phase, the data scientists and stakeholders will have a set of design decisions to set up the technical framework for the project. From this more well-informed perspective, the project scope can be revisited and modified, which may in turn require more data collection, feature engineering, or exploratory analysis.

Example: Drawing on what they learned in exploring the data, the data scientists work with the public health officials to formulate a classification problem at the address level using blood lead levels above a specific level as a training label. Monthly risk scores will be produced for every house with a child younger than 12 months to correspond with the planning cycle of the department's housing inspection team and evaluated on the basis of precision (positive predictive value) among the top 250 highest risk addresses, consistent with their monthly capacity for lead inspections, as well as the representativeness of underserved communities in the results.

Phase VI: Analysis/modeling

Many projects will move through multiple rounds of exploration and refinement, iteratively approaching a final analytical phase as the problem definition and scope continue to evolve. While the competency domains essential to these later analytical phases are similar to those employed in the earlier exploratory work, the specific skills used here will tend to shift away from data description and more toward summarization, prediction, and/or extracting meaning. Generally, this phase involves generating a large number of models, analyses, or results followed by analysis to draw meaningful conclusions. In the case of predictive modeling, this might involve the process of model selection, balancing different performance and fairness metrics to arrive at a single model (or small menu of choices) to put into practice. For analytical projects, this phase may also involve telling a story from the available data, putting to use not only communication and data visualization skills but also ethical frameworks for how to summarize vast amounts of data in fair and meaningful ways.

Example: The data scientists run a grid of thousands of model specifications, including several families of classifiers and hyperparameters. Based on its ability to both achieve high precision in the top 250 and balance false omission rates across race and socioeconomic status, a random forest model was chosen to test in a field trial.

Phase VII: Field validation/pilot

The previous phase results in a final set of analysis results or models that are ready to be piloted or validated in a field trial. This phase of the project involves designing the trial to test the ongoing effectiveness and usability of the analysis. In some cases, this may involve developing a randomized control trial to measure the causal impact of deploying a predictive model, whereas in others it may involve collecting feedback on how a report impacts decision-making. In any case, this phase should focus on validating that the results of the analysis in fact continue to perform as anticipated when presented with truly novel data, including with respect to relevant fairness metrics. Likewise, gathering qualitative feedback from decision makers acting on the analysis is an important aspect of the field pilot.

Example: A 1-year field trial was developed, during which a random 50% of the 250 highest risk addresses were inspected for the presence of lead each month, and remediated where hazards were found. The trial confirmed the performance of the model in identifying children at risk of poisoning because of the presence of lead in their homes as well as its representativeness across communities.

Phase VIII: Taking action

Finally, for a data science project to successfully impact the decisions or actions of policy makers or stakeholders, results must be clearly and effectively communicated to these (generally nontechnical) audiences along with recommended actions or a menu of choices. The ethical obligations of responsible data science practitioners at this stage reach far beyond avoiding the colloquial idea of "lying with statistics" to an awareness of the potential societal impact of their work. Any recommended course of action involves trade-offs (for instance, between optimizing for overall efficiency vs. fairness across affected groups) and the data scientist performing these analyses may be the best-positioned individual to articulate the trade-offs associated with any potential action.

Example: Although the number of households with lead issues remediated was too small to have a significant impact on the number of children diagnosed with lead poisoning during the trial period, calculations suggested that deploying the model could appreciably impact lead poisoning over the following decade. The Department of Public Health decided to move forward with putting it into practice, committing resources to maintain and periodically refresh and reevaluate the model.

Taken alone, each phase of a project draws on a range of different competencies, highlighting the need for well-rounded skill development in data science education. Moreover, the heavily interconnected nature of the project phases illustrated in Figure 1 reflects the importance of agility in applying these skills throughout the entire project lifecycle. Effective data scientists need to be able to recognize when initial exploration or analyses might dictate a change in scope and responsible practitioners will surface new problems they identify in their work with the data as candidates for future work. With their heavy focus on teaching analytical methods or tools through classroom instruction in a linear siloed manner, we feel that many current efforts to educate data scientists fall short of their mandate to produce effective well-rounded practitioners who are equipped to handle the nuances of problems they will encounter in their career.

A Better Way: The Data Science Residency Master's Program

For many of the skill domains described above, classroom-based coursework is an inefficient and ineffective method of building student competency. Instead, we propose a Data Science Residency Master's program structured around a core of applied project work supported by lectures and workshops.

Ideally, projects should be sourced from a diverse set of industry, nonprofit, and government partners and reflect a variety of problem types and methods to give students a range of options to explore and allow them to match with a project that best fits their interests. Projects should be structured to answer a practical question using the partner's actual data and involve the entire scope of project definition through analysis to field testing, to deployment and maintenance. We call them "partners" because they are collaborators in the entire process as opposed to just providing a problem or a data set.

Additionally, we believe that classroom instruction will be most effective if it is tightly coupled with problems students are currently facing in their projects, providing context, frameworks, and tools to help them approach issues that are already at the top of their mind. A number of programs in computer science[29,30] and introductory courses in data science[31] have found advantageous effects to pairing instruction with real-world context through projects, and we have likewise seen this sort of "priming" effect at play several times in working with students in the Data Science for Social Good summer program,[32] informing our perspective on how learning and practice interact. Our proposed approach to instructional hours for Residency Master's therefore breaks from the typical concept of quarter- or semester-long topic-centric courses in favor of a concept of instruction that better reflects the skills students will put to use through the lifecycle of a data science project. This approach builds on the materials we have previously created for the Data Science for Social Good summer program[32,33] as well as for the Coleridge Initiative classes in Applied Data Analytics.[34]

Here, we envision shorter segments of instruction on each topic, closely tied to one another as well as to the project work that forms the core of the student's experience. Each segment could be taught by faculty who specialize in the topic at hand or faculty associated with the data science program itself, but in either case, we believe that it is critical to have both close coordination of classroom curriculum across these domains as well as direct ties between instruction and challenges students are likely to be facing in putting this work into practice.

Our proposed structure is shown in Figure 3. Early in the program, a heavier focus on classroom instruction will ensure all students have a strong foundation in tools and methods, while three "project sprints" will focus on different phases of the data science project lifecycle. In their second and third years, instruction will play a supportive role as students focus primarily on project work, with their responsibilities increasing with seniority, and culminating in an oral defense.

Although typical master's programs are often 2 years in length, we believe the longer program described here provides three benefits: First, as described above, applied projects generally involve a considerable degree of iteration while developing experience with the data and problem domain and both the work itself and students' learning from it are likely to benefit from sufficient time to engage in this process. Second, 3 years provides needed time to develop well-rounded practitioners without overwhelming project work with supplemental instructional time. And, third, we believe that it is valuable for students to have an opportunity to see projects through to field trials and validation on novel data, a process that in itself often takes time. Below we describe our proposed content for each of these 3 years in more detail.

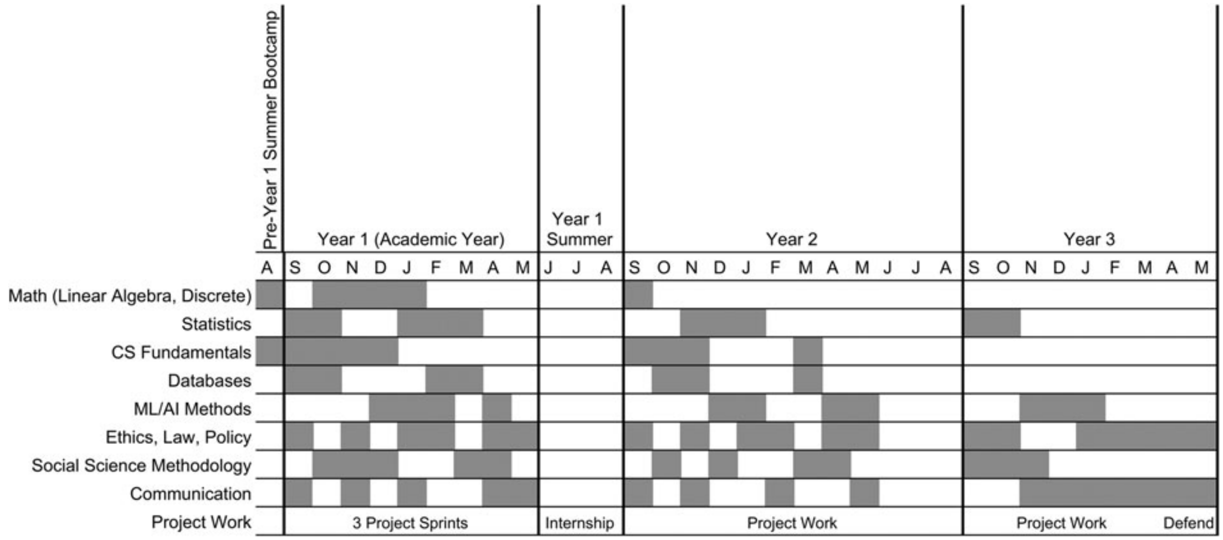


FIG. 3. Proposed distribution of instructional hours and project work for a Data Science Residency Master's program.

First year

Before beginning the first year of their program, students would be expected to participate in (or, optionally, test out of) a boot camp covering mathematics and programming fundamentals to ensure they begin the program with a common foundation. Following this summer introduction, students should be expected to have a basic understanding of linear algebra, probability, SQL, python, and programming workflows (editing and running code, bash, collaboration, and version control, etc.).

The program's first year is also heavier on coursework than the subsequent years to familiarize students with the full data science workflow and begin to develop their competency in each domain. Here, the first year is structured as a microcosm of the entire project lifecycle, supported by three "project sprints" developed from real-world problems and data sets (ideally drawn from the program's previous projects after it has been running for a few years):

- Sprint 1 focuses on the early phases of a project: evaluating need and convincing decision makers to support a project, problem definition and scoping, and data acquisition.
- Sprint 2 focuses on the more technical aspects of exploring the data and performing analyses as well as refining the project's goals and criteria as more knowledge is developed.
- Sprint 3 focuses on consolidating results, model selection, and turning results into action.

Each sprint would run for 3 months with tightly coupled concurrent instruction. For instance, lectures in the domain of ethics, policy, and law during Sprint 1 might focus heavily on the existing legal and ethical practices for acquiring, using, and protecting sensitive data. During Sprint 2, lectures in this domain might focus on understanding and applying metrics to measure bias and fairness. And, in Sprint 3, they could focus on building frameworks for evaluating the potential societal implications of students' work and the inherent trade-offs involved. Likewise, instruction in statistics might evolve from focusing on data exploration early in the program to inference and model interpretability methods later.

During the summer following their first year, students would ideally participate in an internship at the partner organization for the project they will work on in their remaining 2 years. Doing so will allow them to form connections with the project partners as well as develop a working understanding of the organization's work, data, and priorities.

Second year

During the second year, the program's focus shifts to primarily emphasize project work. Classroom instruction should be limited to 3–6 hours per week, generally revisiting topics introduced in the first year with an increased level of depth and nuance. Just as instructional time in first year is coupled to the introductory project sprints, coursework in the second year should seek to provide the tools students will need in their ongoing projects just in time as they need them. Second-year students' project work should begin with scoping and problem definition and shift toward a heavy focus on the technical aspects of understanding and analyzing the partner data after several months.

In our experience, students can also learn a great deal from each other in the course of their project work if given a forum in which to do so. Our experience running project-based programs has shown that several types of interactions can be helpful here: First, frequent (e.g., weekly or biweekly) and highly structured status check-ins across all projects can provide context about how different organizations operate and move through phases of the work. Second, in-depth technical updates (deep dives) from each project on a less frequent basis (e.g., every other month) can provide exposure to methods and applications beyond a given student's project and give students the opportunity to provide feedback on each others' work. And, third, periodic focused skills sessions provide an opportunity to practice applying specific skills to other contexts, such as holding a scope-a-thon once a year to develop ideas for potential new projects based on their current work.

Third year

The role and responsibilities of students should expand in their third year, including mentoring first- and second-year students as well as more direct interactions with project partners, both to communicate results and help develop the scope of future projects. Additionally, the regular cross-project check-ins and technical updates should continue to serve as a forum for exposing students to other problems, organizational contexts, and technical methods beyond their own projects. Formal coursework should again be fairly limited in the third year of the program, with an increased focus on practical curriculum around communicating about data science to nontechnical audiences and understanding the role of data scientists in society.

Finally, we envision an oral defense as the culminating exercise of the student's program, allowing the student to demonstrate their mastery of the data science competencies and their practical application. In particular, we believe that it is critical for the defense committee to include representation from the project partner, giving them a voice in the student's final assessment (and that this expectation should be built into the partner agreement).

Alternative Curriculum Options

Traditional instruction

While we believe the structure proposed above would best suit learning and retention by matching classroom instruction to the challenges students face at each phase of their project work, we also recognize the departure from traditional single-instructor, single-topic courses may be challenging to implement in some contexts. As such, Table 2 describes a structure that maintains the core elements of the Data Science Residency Master's but may fit more readily into existing settings.

While this version of the program uses a more traditional structure for the coursework, we nevertheless believe that it is important for this time in the classroom to be envisioned as primarily a means to support students' success in their practical work through real-world projects. As above, the first year pairs a heavier focus on classroom instruction with three project sprints to introduce students to tools, methods, and theory they will need to apply to successfully, and responsibly, execute a data science project. Ideally, the sprints could be integrated with the assignments across these core courses. The second and third years allow students to focus more heavily on their project work with a considerably reduced course load, and the degree program culminates in an oral defense.

Table 2. Alternative structure for Data Science Residency Master's program

Term	*Coursework*	*Project work*
Pre-year 1 summer	Mathematics/Programming (Python and SQL)/Workflow Boot camp	
Year 1	Math (Linear Algebra and Discrete Math), Statistics, Computer Science Fundamentals (Data Structures and Algorithms), Databases, Machine Learning Methods, Social Science	Three project sprints
Year 1 summer		Internship with project partner
Year 2	Advanced Methods (Machine Learning, Causal Inference), Ethics/Fairness, one Elective	Project work
Year 3	Communication, Advanced Methods Elective (from Machine Learning, Stats, Operations Research, Social Sciences), one Elective	Project work, defense

Two-year structure

The 3-year structure described above is intended to provide a sufficient amount of time to develop both a broad foundation and an analytical depth through applied work to help students develop as well-rounded practitioners. However, we recognize that this represents a departure from a more typical 2-year master's program and imagine that some institutions may be interested in integrating some of the concepts described here in a shorter structure. Even in an abbreviated 2-year structure, project work should serve as the core component of the program. A suggested structure might include:

- The initial summer boot camp and first year similar to the first year of the 3-year program described above, however, with the project sprints shortened to 1 month each allowing students to begin the initial phases of scoping and data exploration with their project partner in the spring term.
- The summer internship between the first and second years as a means of fostering trust, relationships, and domain knowledge for the ongoing project work.
- A second year similar in structure to the third year of the program above, heavily focused on project work (including routine check-ins and technical updates) with at most the equivalent of one to two courses of instructional time per term and culminating in an oral defense.

Staffing Structure

The real-world project work and external partnerships that are the central components of the Data Science Residency Master's have a number of implications for the structure and staffing the program will need to be successful.

First and foremost, the program will need to develop a pipeline of new projects through partnerships with government agencies, industry, and nonprofits and will likely need staff who can cultivate and maintain these relationships. Maintaining a small team size for each project will be important to ensuring a well-rounded experience in which students participate in all aspects of the work (in practice, our experience with the Data Science for Social Good program might suggest that groups of more than three to four students can result in siloing). As such, the total number of projects will be an important determinant of the enrollment capacity of the program.

These external partnerships are critical to the success of the program overall. A good project partner is far more than a source of data for students to work with independently. Rather, they are an integral part of the project from end-to-end, committing resources and time to project scoping, extracting and explaining data, providing intermediate feedback and "gut checks," potentially executing on field trials, and participating in students' final evaluation. Moreover, most external partners will have several problems that can be developed into data science projects (and, often, many more to be discovered through the course of this work), so external partnerships should be envisioned as long-term relationships well worth investing programmatic resources in cultivating.

Second, the project work will need to be supported with human and technical infrastructure. This includes project managers who can help create and facilitate the ongoing relationships with the external partners through the course of the project as well as professionally managed servers and databases by a staff of DevOps engineers. These staff members can likely work across several projects and external partners, but the program will need to ensure both roles are staffed sufficiently to allow the project work to progress smoothly.

Third, a staff of full-time practicing data scientists could act as project leads and technical mentors, as well as providing a much-needed source of institutional knowledge about projects, tools, and data sets as students enter and leave the program. While institutional structure will dictate the specifics of the role (whether faculty, research faculty, senior fellow, or staff scientist), we envision it as a permanent position supported by hard money from the degree program and focused full time on mentorship and supporting the applied data science work of the projects through practical research and tool development.

Specialized Versions of the Master's Program

Structurally, the Data Science Residency Master's would ideally be envisioned as a collaboration that bridges several departments at an institution: computer science, statistics, public policy, government, medicine, public health, and potentially many others. While we recognize that creating a new cross-institution entity may pose a challenge for some institutions, there are several strategies and opportunities to foster broad participation and inclusion across existing departments and faculty.[16,35–37]

Existing data science programs might also consider adding the Residency Master's as an optional track or degree alongside their current curriculum. In addition to providing a smoother path to transition to a more clinical program over time, this could provide a good opportunity to study the outcomes of students who complete the two tracks to further refine the model and understand what additional skills they may need.

Similarly, existing Masters in Public Policy, Public Administration, Social Work, and Public Health programs can add components of this model as a certificate or specialization to prepare professionals in those disciplines. These specialized programs would have the curriculum be more expansive at the intersection of computer and social sciences, and the projects centered around real ongoing challenges faced by the public sector that can benefit from the use of data science. The partners here might be governments, nonprofits, foundations, media organizations, or even private companies. We are currently exploring further developing this type of program that would allow us to train effective public sector data scientists and augmenting existing programs in Public Policy, Public Administration, and Social Work. We believe that the creation of the recently Public Interest Technology University Network[38] is timely, and not only reflective of the need for such a program but also a good venue for collaboration across universities as these programs get created.

Discussion

Current data science degree programs, with their heavy focus on coursework and theory, offer students little opportunity and training to undertake a significant project that reflects the messiness and constraints of working in a professional context. Viewed through the lens of competency domains that need to be mastered by effective practitioners, it becomes apparent that this current educational paradigm leaves significant gaps to be filled (one hopes) by their early years of experience in the workforce.

Here, we have proposed a new approach to training data scientists, taking a cue from the concept of medical residency and putting the development of practical experience at the center of a student's education. This approach has been informed by our experience with designing and implementing different types of data science education programs including the Masters in Computational Analysis and Public Policy at the University of Chicago,[25] the Data Science for Social Good Summer Fellowship,[32] and the Applied Data Analytics for Governments program at the Coleridge Initiative.[39] We believe that this focus on learning by doing will develop more well-rounded data science graduates who are better equipped to begin their professional careers. At present, we are exploring opportunities and collaborators to test this model and encourage others to do the same, as well as experiment with variations on the program proposed here.

An important aspect of future implementations of this or any other novel approach to training data science practitioners is improving measurement of the performance of these educational efforts. In collecting our thoughts here, based both on our own experiences and literature highlighting some of the limitations of current approaches, we have found little evidence to assess the potential social impact of different training approaches in a quantitative manner. Pairing steps toward new approaches with steps toward better measurement would surely benefit both programs and students alike (of course, as individuals who hope to see increasing numbers of data science practitioners in the public sector, we would also urge the field to seek out a more holistic view of degree program success than statistics focused compensation). As the discipline of data science continues to develop, much remains to be learned about best practices in training and education. We hope that the degree program model we have proposed will not only provide one step forward along that path of discovery but also spark other forms of innovation and experimentation aimed at training our students to have a positive practical impact through data science.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

No funding was received for this article.

References

1. Long DM. Competency-based residency training. Acad Med. 2000;75: 1178–1183.
2. Sanchez C. Programs train teachers using medical school model. NPR morning edition. 2010. Available online at https://www.npr.org/templates/story/story.php?storyId=125854975 (last accessed August 6, 2019).
3. Arizona Teacher Residency Model Gets High Marks. US Dept. of Ed. Office of innovation and improvement. 2014. Available online at https://sites.ed.gov/oii/2014/06/arizona-teacher-residency-model-gets-high-marks (last accessed August 6, 2019).
4. Baumer B. A data science course for undergraduates: Thinking with data. Am Statistician. 2015;69:334–342.
5. Hicks SC, Irizarry RA. A guide to teaching data science. Am Statistician. 2018;72:382–391.
6. Horton NJ, Baumer BS, Wickham H. Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics. arXiv. 2015;arXiv:1502.00318.
7. National Research Council. Training students to extract value from big data: Summary of a workshop. Washington, DC: The National Academies Press. 2015.
8. Cao L. Data science education. In: Cao L, Yu PS (Eds.): Data science thinking. Cham, Switzerland: Springer International Publishing. 2018. pp. 329–348.
9. Grimshaw SD. A framework for infusing authentic data experiences within statistics courses. Am Statistician. 2015;69:307–314.
10. Bailer J, Hoerl R, Madigan D, et al. Report of the ASA workgroup on master's degrees. American Statistical Association. 2012. Available online at https://magazine.amstat.org/wp-content/uploads/2013an/masterworkgroup.pdf (last accessed August 6, 2019).
11. Hardin J, Hoerl R, Horton NJ, et al. Data science in statistics curricula: Preparing students to 'think with data'. Am Statistician. 2015;69:343–353.
12. Cobb G. Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. Am Statistician. 2015; 69:266–282.
13. Brown EN, Kass RE. What is statistics? Am Statistician. 2009;63:105–110.
14. Hardin J. Dynamic data in the statistics classroom. arXiv. 2016;arXiv: 1603.04912.
15. Dichev C, Dicheva D, Cassel L, et al. Preparing all students for the data-driven world. Winston Salem, USA: ADMI, March 31–April 3, 2016.
16. Song I-Y, Zhu Y. Big data and data science: What should we teach? Expert Syst. 2016;33:364–373.
17. Starr P. The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry. New York, NY: Basic Books. 1982.
18. Patil D, Mason H, Loukides MK. Ethics and data science. Sebastopol, CA: O'Reilly Media. 2018.
19. Floridi L, Taddeo M. What is data ethics? Philos Trans A Math Phys Eng Sci. 2016;374:20160360.
20. Hajian S, Bonchi F, Castillo C. Algorithmic bias: From discrimination discovery to fairness-aware data mining. San Francisco, CA: KDD, August 13–17, 2016. doi:10.1145/2939672.2945386.
21. Meeting #5: Integrating ethics and privacy concerns into data science education. NAS roundtable on data science postsecondary education. 2018. Available online at http://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_185108.pdf (last accessed August 6, 2019).
22. Master of Science in Analytics. Northwestern University. 2019. Available online at https://www.mccormick.northwestern.edu/analytics (last accessed August 6, 2019).
23. Meeting #4: Alternative institutional and educational mechanisms. NAS roundtable on data science postsecondary education. 2017. Available online at http://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_183475.pdf (last accessed August 6, 2019).
24. Halicioglu Data Science Institute. UC San Diego. 2019. Available online at https://datascience.ucsd.edu (last accessed August 6, 2019).
25. Master of Science in Computational Analysis and Public Policy. University of Chicago. 2019. Available online at https://capp.uchicago.edu (last accessed August 6, 2019).
26. Joint PhD Program in Machine Learning and Public Policy. Carnegie Mellon University. 2019. Available online at https://www.ml.cmu.edu/academics/joint-phd-mlpp.html (last accessed December 3, 2019).
27. Rodolfa E, Bent R, Eisman E, et al. A cube model for competency development: implications for psychology educators and regulators. Prof Psychol Res Pract. 2005;36:347–354.

28. Data Science Project Scoping Guide. Center for Data Science and Public Policy, University of Chicago. 2018. Available online at https://dsapp.uchicago.edu/home/resources/data-science-project-scoping-guide (last accessed November 20, 2019).
29. Saltz J, Serva MA, Heckman R. The GET immersion experience: A new model for leveraging the synergies between industry and academia. J Inf Syst Educ. 2013;24:121–131.
30. Grisham PS, Krasner H, Perry DE. Data engineering education with real-world projects. ACM SIGCSE Bull. 2006;38:64.
31. Saltz J, Heckman R. Big Data science education: A case study of a project-focused introductory course. Themes in Sci Technol Educ. 2015;8:85–94.
32. Data Science for Social Good Fellowship. 2019. Available online at www.dssgfellowship.org (last accessed December 3, 2019).
33. Hitchhiker's Guide to Data Science for Social Good. 2019. Available online at https://dssg.github.io/hitchhikers-guide (last accessed December 3, 2019).
34. Jupyter Notebooks for Applied Data Analytics. Coleridge Initiative. 2019. Available online at https://github.com/Coleridge-Initiative (last accessed October 17, 2019).
35. National Academies of Sciences, Engineering, and Medicine. Data science for undergraduates: opportunities and options. Washington, DC: The National Academies Press. 2019.
36. Williams T. Diversity and inclusion in data science: Using data-informed decisions to drive student success. Webinar presentation to the committee on envisioning the data science discipline: The undergraduate perspective. 2017. Available online at www.nas.edu/envisioningDS (last accessed August 6, 2019).
37. Rawlings-Goss R. Keeping data science broad: Negotiating the digital and data divide among higher education institutions. South big data innovation hub. 2018. Available online at http://bit.ly/KeepingDataScienceBroad_Report (last accessed August 6, 2019).
38. Public Interest Technology University Network. 2019. Available online at https://www.newamerica.org/public-interest-technology/university-network/about-pitun (last accessed October 17, 2019).
39. Coleridge Initiative. 2019. Available online at http://coleridgeinitiative.org (last accessed October 17, 2019).

Cite this article as: Rodolfa KT, De Unanue A, Gee M, Ghani R (2019) An experience-centered approach to training effective data scientists. *Big Data* 7:4, 249–261, DOI: 10.1089/big.2019.0100.