Unleashing the Problem-Solving Potential of Next-Generation Data Scientists

Lizhen Zhu and James Z. Wang

Abstract Data science, an emerging multidisciplinary field, resides at the intersection of computational sciences, statistical modeling, and domain-specific sciences. The current norm for data science education predominantly focuses on graduate programs, which presume a preexisting knowledge base in one or more relevant sciences. However, this framework often overlooks those who don't plan to pursue graduate studies, thereby limiting their exposure to this rapidly expanding field. Penn State addressed this gap by establishing one of the first undergraduate degree programs in data sciences, a collaboration between the College of Information Sciences and Technology, the Department of Computer Science and Engineering, and the Department of Statistics. One key component of this program is a project-focused, writing-intensive course designed for upper-class undergraduates. This course guides students through the entire data science project pipeline, from problem identification to solution presentation. It allows students to apply foundational data science principles to real-world problems, advancing their understanding through practical application. This chapter details the objectives, rationale, and course design, alongside reflections from our teaching experience. The insights provided could be helpful to instructors developing similar data science programs or courses at an undergraduate level, broadening the influence of this important field.

Keywords Undergraduate data science education · Project-based learning · Interdisciplinary curriculum · Interpersonal skills · Intrapersonal skills

Introduction

With the rapid advancements of digital sensing and database technologies, in today's world, data is ubiquitous, and its significance cannot be overemphasized. Data isn't

L. Zhu · J. Z. Wang (✉)
College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
e-mail: ljz5180@psu.edu; jwang@ist.psu.edu

J. M. Carroll (ed.), *Innovative Practices in Teaching Information Sciences and Technology*, https://doi.org/10.1007/978-3-031-61290-9_8

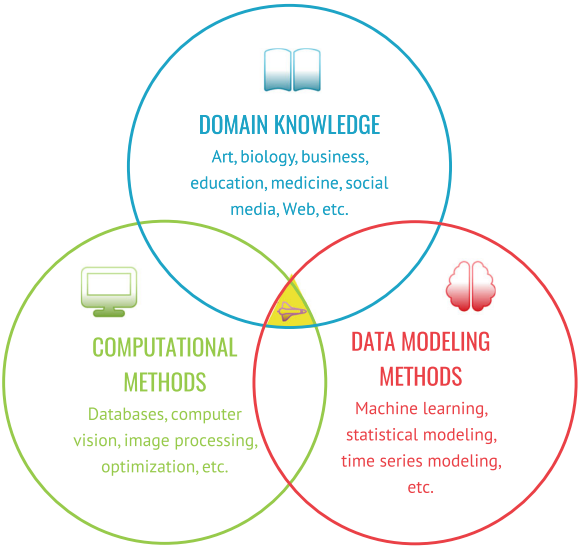

Fig. 1 Data science projects require domain knowledge, computational methods, and data modeling methods

just an integral part of our daily lives but is also used by scientists to discover new knowledge. A key challenge in leveraging the growing availability of data, however, lies in extracting meaningful insights from the ever-growing sea of information.

Data science is an emerging multidisciplinary field, crossing computational science, mathematical and statistical sciences, and application domain sciences that serves as a solution to this challenge. As depicted in Fig. 1, domain knowledge encompasses the understanding of the field of application, computational methods refer to the algorithms and processes used for data analysis, and data modeling methods consist of techniques for constructing and fine-tuning models. The intersection of these three components is where data science operates most effectively, leveraging domain expertise to apply appropriate computational and modeling methods, thereby driving valuable insights from data.

The demand for professionals with data science expertise has skyrocketed. Jobs in this sector are projected to increase by 36% from 2021 to 2031 (U.S. Bureau of Labor Statistics, 2023). However, because data science has to build on the foundational knowledge of these core fields, most training programs are at the graduate level, leaving undergraduate students without an opportunity to explore this exciting and rapidly expanding field.

At Penn State, we recognized this gap and created one of the first undergraduate data science degree programs in 2016 as an intercollege program between the College of Information Sciences and Technology, the Department of Computer Science and Engineering of the College of Engineering, and the Department of Statistics of the Eberly College of Science (Hallman, 2018). The data science major is structured into three options: computational, statistical modeling, and applied data sciences, each catering to different potential career trajectories.

The computational option emphasizes core computational skills, such as the design, implementation, and analysis of software that manages the volume, heterogeneity, and dynamic characteristics of large datasets and that leverages the computational power of modern hardware such as CPU and/or GPU cluster computers. The statistical modeling option focuses on statistical models and methods that are foundational to discovering and validating patterns in big data. The applied option equips students with the skills to integrate knowledge and techniques in problem-solving. Specifically, the emphasis is on addressing pressing scientific, organizational, and societal challenges.

Although students of the three options take different courses in their first two years, covering core methodologies relevant to their option, we begin to bring them into real-world problem-solving using data science in the third year through a required course, "applied data sciences," developed by James Wang, a coauthor of this chapter. Wang's diverse educational background includes advanced degrees in mathematics, computer science, and medical information sciences. His extensive experience in data science research spans a wide array of fields, including art, healthcare and medicine, plant biology, psychology, and meteorology.

This cornerstone course includes students from all three options and enables them to apply data science principles to real-world issues while sharpening their writing skills. This course is designed to provide students with a comprehensive understanding of data science by introducing fundamental principles through real-world examples and demonstrating how these principles apply to various methods and techniques commonly covered in data science courses. The course includes a term project and public speaking opportunities to enhance students' learning experiences. Upon completion, students should be able to proficiently navigate the entire process of a data science project, from problem formulation to data collection, technique selection, implementation, and performance evaluation. They will also be adept at effectively communicating actionable insights through written reports and oral presentations.

In this chapter, we will share our experiences in developing and teaching this project-focused, writing-intensive course. We will discuss the course design, the rationales behind its structure, and the insights we have gained from teaching it. We hope that our experiences could serve as a guide for instructors from other undergraduate data science programs looking to develop similar courses or programs.

Course Design

The course is structured to equip undergraduate students with the necessary knowledge and skills for executing data-driven projects in professional or research settings. After taking this course, students should master a comprehensive understanding of the data science pipeline. As illustrated in Fig. 2, a typical pipeline involves (1) problem formulation; (2) data acquisition; (3) data preparation, clean-

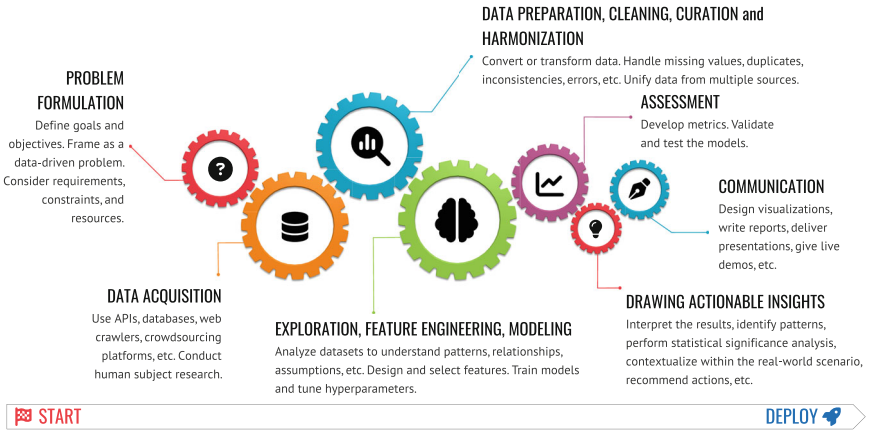

Fig. 2 Typical data science project pipeline. This diagram illustrates the sequential steps in a data science project, from problem formulation to communicating data science solutions to stakeholders

ing, curation, and harmonization; (4) exploration, feature engineering, and modeling; (5) assessment; (6) drawing actionable insights; and (7) communication.

To achieve this objective, the course is segmented into three key parts:

- **Foundation for Carrying Out Large-Scale Data Science Projects:** We include the concept of a data science project as an integration of computational methods, mathematical or statistical methods, and domain knowledge (Fig. 1). To motivate students about the importance of data science, we emphasize the view of data science as the fourth paradigm of conducting scientific exploration, following empirical science, theoretical science, and computational science (Fig. 3). Students are taught commonly used tools and are also equipped with nontechnical but vital skills like academic writing and presentation. Brief lectures from the instructor and teaching assistant (TA) on these topics are followed by simple in-class and take-home assignments to assess students' understanding. Furthermore, an in-depth analysis of the pipeline through a case study is used to reinforce knowledge.
- **Exposure to Important Data Science Subfields:** Based on their individual interests, students are assigned various topics to explore and present a talk summarizing the related literature. Seed papers provided by the instruction team aid the presenter in gathering additional materials. Basic literature search skills such as using various online databases, e.g., Google Scholar, IEEE Xplore, and ACM Digital Library, are covered by the instructor. During the student presentation, the presenter leads the discussion. The talk is peer-reviewed according to set criteria. Presenters are encouraged to promote audience participation. A subset of the seed papers also serves as reading assignments for all students. To ensure that students are keeping up with their reading, readiness assessment tests (RATs) are conducted in class based on the assigned reading. In addition

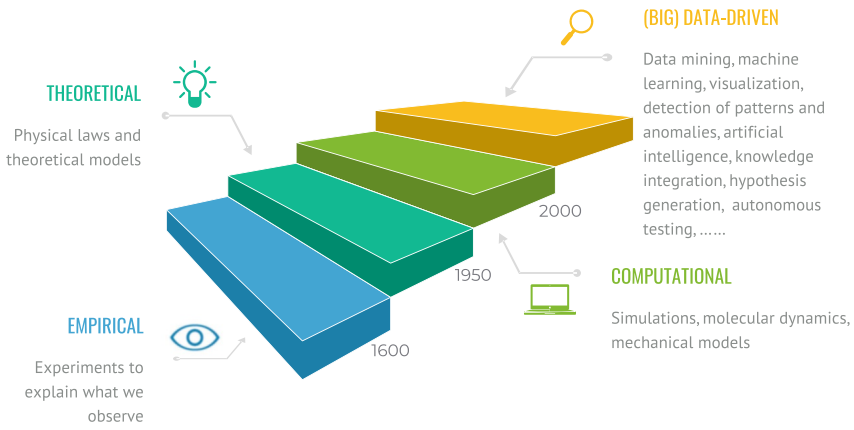

Fig. 3 The primary paradigms of scientific inquiry: empirical, theoretical, computational, and (big) data driven. Empirical science is rooted in observation and experimental data. Theoretical science focuses on the development of models and theories to explain phenomena. Computational science uses mathematical models and algorithms to solve complex problems. Finally, data-driven science utilizes sophisticated techniques to extract knowledge and insights directly from structured, semi-structured, and unstructured data

to this formal presentation assignment, the course also offers an opportunity for students to give an informal presentation on an interesting data science-related topic. The objective is to keep students up to date on the latest and most exciting developments in the field.

- **Real-World, Team-Based Project:** Students, in groups of two or three, work on a problem with real-world relevance. The problem, selected at the start of the semester in consultation with the course instructor and TA, should aim to utilize a large volume of complex digital data and statistical data modeling methods. Each team prepares a short one-to-two-page project proposal and an ignite talk (i.e., short, fast-paced presentation where the team shares their project ideas and approaches) at the beginning of the semester and receives writing and technical feedback for further enhancement. Regular meetings with the instructor and TA are set up to discuss project progress and provide detailed feedback on their reports. Toward the end of the semester, each team presents their project for peer review and submits an 8-to-10-page project report in a scholarly publication format for grading.

Student grades are composed of three elements.

1. **Participation** (30%): This part includes random class attendance checks and in-class activities, accounting for 15% of the grade. The other 15% is based on RATs. Over the semester, six RATs are administered, with the lowest score being disregarded.

2. **Presentations** (20%): Students will present on recent research and developments in the field, which will make up 20% of the final grade. The presentations will be peer-reviewed, taking into account their breadth, depth, clarity, and style.
3. **Project** (50%): The term project is evaluated based on its applicability, correctness, design, and functionality. The proposal and two midterm reports are worth 3%, 5%, and 7% to the grade, respectively. The final report, submitted in a scholarly article format, accounts for 20% of the grade. The last 15% comes from the final in-class presentation and demonstration of the project's functionalities.

The course nurtures multiple competencies essential for data scientists, including (1) technical, (2) intrapersonal, (3) interpersonal, and (4) problem-solving skills (Fig. 4). A recent survey of industry experts reaffirms the importance of these competencies (Weiser et al., 2022). In the following sections of the chapter, we will provide details on how we have structured the course to develop each of these skills in our students.

Fig. 4 An overview of the diverse skills targeted in the course curriculum, spanning technical, interpersonal, intrapersonal, and problem-solving capabilities, designed to equip students for success in their professional data science careers. Some of the skills can be categorized into multiple areas

Technical Skills: Continuous Learning, Case Analyses

This course provides an extension of the technical knowledge acquired in previous courses. There are mandatory prerequisites, including computer programming, calculus, probability, machine learning, data management, and privacy and security. Visual analytics is a suggested preparation. While previous experience with complex data types (e.g., multimodal/multimedia data, digital images, graphs) can be beneficial, it is not a requirement. During the initial phase of the course, we supplement students' existing knowledge through lectures on carefully chosen topics. These include the Message Passing Interface (MPI), PyTorch, cluster computing, cloud computing, and time series analysis.

Because data science is rapidly evolving with daily updates in methodologies and tools, successful data scientists require not just robust technical skills but also the ability to learn independently. Students need to develop the ability to source, comprehend, and use online scholarly articles, tutorials, textbooks, and short courses effectively. Furthermore, refining skills such as summarizing, explaining, discussing, presenting, critiquing, and defending scholarly work is an important aspect of their development in this field.

Continuous Learning In our pedagogical approach, we leverage *continuous learning* as a vehicle to not only reinforce students' cognitive grasp of the material but, more crucially, to provide them the opportunity to learn to learn, thereby equipping them with the adaptability and innovative thinking necessary for navigating future environmental shifts (Business Information Review Editorial, 2018). Continuous learning is characterized as a dynamic, unceasing cycle necessitating continuous modification and adaptation in response to feedback. Figure 5 depicts the cyclical process, which includes classroom instruction, assignments, and feedback.

The cycle begins with "classroom instruction," during which students acquire and process information, laying the groundwork for their knowledge base. The subsequent stage, "assignments," involves the integration and application of learned material, deepening comprehension, and sharpening skills. Finally, the "feedback" stage loops back into the cycle, providing critical insight into students' areas of strength and opportunities for growth. This reflective stage facilitates the refinement of both knowledge and skills, preparing the students for the next round of instruction and learning. The overarching objective of this methodology is the **holistic** development of students, promoting not only academic proficiency but also a lifelong passion for learning and self-improvement, critical attributes for a career in data science.

Each segment of this pedagogical cycle can adopt various formats. Assignments can take the form of in-class practical exercises, post-class homework, and presentations. In-class hands-on activities encourage students to replicate or make minor modifications to classroom demonstrations. This process empowers students to surmount perceived obstacles and fosters a willingness to experiment with novel concepts, such as coding with new software packages or formatting documents in unfamiliar LaTeX templates. Instructors can observe students during

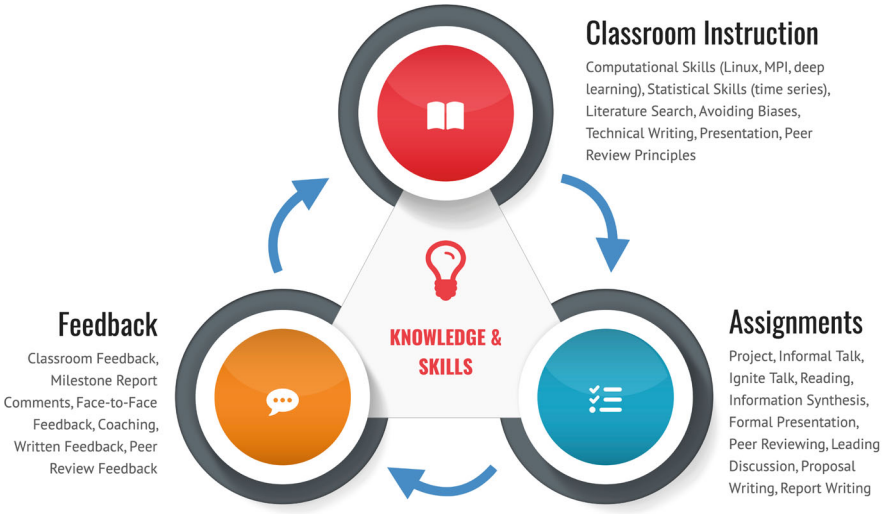

Fig. 5 This diagram encapsulates the cyclic and progressive nature of the learning process

these exercises, providing guidance to those facing difficulties and identifying common areas of misunderstanding for emphasis in subsequent instruction. Peer-to-peer interaction can also be encouraged to offer another form of feedback.

Homework serves as an opportunity for students to deepen their comprehension and requires more effort than in-class activities. Students are expected to engage in reflective thought or research to complete these assignments. As a form of feedback, grades alone may be insufficient. A feedback summary outlining common errors could be disseminated; even if some students do not commit these errors, it does not necessarily indicate comprehension. Additionally, brief, personalized comments can highlight specific mistakes in each student's work.

The design of subsequent course content can be influenced by students' demonstrated abilities. For student presentations, a peer-review process is often the most effective. The final project is crafted around this repetitive learning cycle, ensuring that our students remain on a path of consistent exploration and development.

Case Studies One crucial element we teach students in this course is the proper construction of a data science project pipeline, as depicted in Fig. 2. We introduce a case study at the semester's beginning to demonstrate the various stages involved in this process.

This case study revolves around the collaborative work of researchers (Li et al., 2012), spanning the fields of computing, statistics, and art history. This multidisciplinary team developed an innovative, data-driven approach to analyze Vincent van Gogh's distinct brushstroke styles. Through this research, they undertook a detailed statistical analysis of a broad set of features from automatically extracted brushstrokes. This analysis compared van Gogh's stylistic evolution over

Fig. 6 Automatic brushstroke extraction for *A Pair of Leather Clogs* (Saint-Rémy-de-Provence, Autumn 1889, oil on canvas, 32.2 cm × 40.5 cm) by Vincent van Gogh (Li et al., 2012). Painting image courtesy of the Van Gogh Museum, Amsterdam (Vincent van Gogh Foundation). The brushstroke map is provided by the James Z. Wang Research Group (The Pennsylvania State University)

different periods and how it contrasts with the styles of his contemporaries. The team developed a new brushstroke extraction technique, fusing edge detection and clustering-based segmentation processes, to handle very high-resolution images with a large number of objects (i.e., brushstrokes). Figure 6 shows the automated brushstroke extraction result for one of the paintings.

The study's results revealed a pronounced rhythmic quality in van Gogh's brushstrokes. That is, van Gogh employed consistently shaped brushstrokes arranged closely together, thereby creating a patterned and recurrent impression. Interestingly, the characteristics that set apart van Gogh's paintings throughout various stages of his career are different from those that differentiate his works from those of his contemporaries.

We've chosen this as a notable case study due to its four distinctive characteristics:

- **Domain Knowledge:** The field of application, art history, is often perceived as separate from the disciplines of computing or mathematics. The success of this study underlines the significant role data science can play in addressing complex, real-world problems. It further illustrates the potential enhancements a data science project can gain from incorporating domain knowledge and closely collaborating with domain experts.
- **Bias Management:** The research highlights the crucial task of bias management at different stages of the data science pipeline. More discussions about bias in visual artwork analysis can be found in a recent scholarly paper (Zhang et al., 2022).
- **Innovative Approaches:** The solution to the problem required the invention of new methods rather than simply integrating existing tools. In particular, we stress to students the importance of appropriate problem-solving approaches when dealing with limited data.

- **Explainability:** In certain fields like art history and medicine, the capacity to elucidate the reasoning behind a model's decisions often outweighs the importance of numerical performance results. This work introduces a pipeline that yields entirely interpretable outcomes.

Throughout the semester, we incorporate a variety of recent data science research papers into the curriculum as reading assignments. This allows students to familiarize themselves with contemporary techniques, methodologies, and applications within the field. Table 1 lists the topics and example papers we assign. We will further elaborate on this aspect of the course in a subsequent section.

Table 1 Reading list for the course, with key topics and associated paper references. NLP: Natural language processing. AI: Artificial intelligence

Field	Topic	Papers
Machine learning	Unsupervised learning, self-supervised learning	Kingma and Welling (2014) and Chen et al. (2020)
	Few-shot learning, meta learning	Koch et al. (2015) and Finn et al. (2017)
	Reinforcement learning	Mnih et al. (2015)
	Adversarial machine learning	Moosavi-Dezfooli et al. (2016)
	Explainable AI	Lundberg and Lee (2017)
	Knowledge distillation	Hinton et al. (2015)
	Domain adaptation	Pan et al. (2010)
Computer vision	Object detection, semantic segmentation	Long et al. (2015) and Ren et al. (2015)
	Image generation, inpainting, style transfer	Gatys (2016) and Zhu et al. (2017)
	Human pose estimation, action recognition, facial recognition, head reconstruction, person reidentification	Schroff et al. (2015) and Cao et al. (2021)
NLP	Machine translation, sequence to sequence	Sutskever et al. (2014) and Cho et al. (2014)
	Language model	Vaswani et al. (2017) and Devlin et al. (2019)
Other applications	Image esthetics, computational photography, computer graphics, augmented reality	Datta et al. (2006) and Mildenhall et al. (2021)
	Recommendation system, music, speech, anomaly detection, fraud detection, robotics and autonomous driving	Sarwar et al. (2001) and Cadena et al. (2016)
	Healthcare, biology, astronomy, meteorology	Jumper et al. (2021) and Chen et al. (2014)
	Internet of things, finance, sports, privacy, law	Aletras et al. (2016) and Li et al. (2018)

Intrapersonal Skills: Self-Reflection, Personal Growth

The acquisition of relevant technical skills represents merely one facet of the knowledge needed for success in the field of data science. Equally critical are intrapersonal skills, comprising the individual's aptitude to comprehend and regulate their own emotional, motivational, and behavioral states. The development of these skills is closely tied to a student's background, leading to significant gaps in intrapersonal skill development across our highly diverse student population. Compounding this issue, some students may lack sufficient awareness of the importance of such skills. Against this challenging backdrop, we prioritize the fostering of intrapersonal skills that are directly related to problem-solving, facilitated through highly flexible task configurations.

Personal Motivation The inevitability of challenges in problem-solving underscores the need for personal motivation. This quality helps students set definitive goals, exhibit perseverance and resilience, and advance through self-guided learning. We use questionnaires to gather students' backgrounds and interests prior to course commencement. All topic choices for presentations, literature reviews, and projects are essentially student driven, reflecting their individual interests. This autonomy, however, presents a unique challenge: instructors must provide tailored guidance for each unique topic and ensure fair evaluation of assignments, irrespective of their subjects.

Self-Confidence Self-confidence is integral to the future success of data science students. Confidence emboldens students to assume risks and cultivate a positive self-perception, which is beneficial to their long-term well-being. We create a dynamic classroom environment and incorporate several presentation assignments coupled with question-and-answer sessions to encourage communication and opinion sharing. Our feedback approach is balanced, identifying areas of improvement while acknowledging and appreciating the students' efforts and progress.

Time Management and Organization Given the academic rigor faced by data science students, having to learn multiple interrelated fields, effective time management and organizational skills are paramount. Despite occasional student complaints about inadequate time to complete assignments and requests for deadline extensions, we strive to align assignment completion times with credit weighting. For extensive projects and literature review presentations, students' time management proficiency significantly impacts their performance. We emphasize the importance of effective time allocation at the project inception and maintain two milestone reports to monitor progress and provide timely reminders. To uphold the focus on learning, we accommodate assignment extensions, reinforcing the importance of thoroughness and quality in completed work. However, to ensure fairness, late submissions are subject to penalties.

Self-Reflection The capacity for self-reflection facilitates continuous progress. Students, through critical assessment of their strengths, weaknesses, and areas of

improvement, can recalibrate their approach and foster ongoing development. This practice necessitates a higher degree of self-awareness and encourages an iterative, improvement-focused learning process. We ask students to include a thorough discussion of the merits and limitations of their approach in their project reports and presentations. Additionally, we advise them to remain cognizant of the various assumptions incorporated in their pipeline and to understand instances where such assumptions may not hold in the real world.

Interpersonal Skills: Effective Communication, Team Collaboration

In their daily work, data scientists frequently interact with a diverse array of stakeholders, including team members, domain experts, clients, senior management, and financial backers. Proficient interpersonal skills not only facilitate effective presentation of their findings and perspectives but also foster an enhanced understanding of other stakeholders' views, thereby sparking further exploration of specific topics.

Understanding Others The first step toward fruitful communication is to understand the perspectives of stakeholders from various fields and backgrounds. These stakeholders bring diverse views, methodologies, and feelings. In the term project, we strive to assign students from diverse backgrounds to the same team. This approach allows them to encounter varying viewpoints and learn to harness such diversity to their advantage. In our RATs, answers to most questions aren't directly available in the papers students read. Instead, they require students to grasp the principal ideas within these papers. Questions may include describing a method's pipeline or comparing the pros and cons of multiple methods. During student presentations, students in the audience are urged to engage by asking and answering questions. This way, they practice understanding each other through careful listening and resolve issues by reaching a consensus.

Being Persuasive Data science is not a field where mathematical proof is the norm. Yet, its methods must be useful in real-world scenarios. A data scientist needs to convey viewpoints clearly and understandably so others can appreciate their work. Adopting empathy, or putting oneself in the shoes of others, can make one's stance more acceptable. Being persuasive is, therefore, critical for data scientists. It enables them to communicate their findings and rationales effectively, secure support for their projects, and bridge the divide between technical and nontechnical audiences. This course helps develop this skill through the term project, where students work collaboratively with teammates to persuade the audience, through both writing and peer-reviewed talks that their problem formulation, approaches, and results are convincing.

Academic Writing Because this course is designated a "writing across the curriculum" course at Penn State, students are provided frequent and significant opportunities to write, revise, and discuss their writing. We introduce essential guidelines for proper formatting, citation, and plagiarism avoidance at the beginning of the course. We developed comprehensive grading rubrics to ensure clarity and fairness in evaluation. By incorporating the cycle illustrated in Fig. 5, we place an emphasis on written communication, aiming to ensure that student-generated reports are both comprehensible and professional. However, improving writing is a gradual process. Despite persistent feedback and correction, common issues such as citation format errors, grammatical and spelling mistakes, and the use of unprofessional visuals continue to be prevalent.

AI-Assisted Writing In the contemporary educational landscape, AI-based writing tools like ChatGPT and Grammarly have gained widespread popularity. While they can be a valuable asset in enhancing language fluency and readability, educators globally grapple with their proper use in writing-intensive courses. The foundational assumption remains—true learning necessitates comprehensive practice. This axiom, though rooted in the pre-digital age, retains its relevance even in today's AI-dominated era.

Notwithstanding, these AI tools have a unique advantage—they provide real-time feedback, which, when effectively utilized, can expedite the learning process. As a result, our course policy encourages students to use these tools, albeit with a caveat—the tools should be used strictly for enhancing language accuracy and readability and *not* for initial content generation. We underscore the significance of students developing their own writing skills and cultivating a distinctive "voice." Additionally, we mandate students to disclose the use of such AI assistance in their submitted writing assignments.

Oral Presentation We also develop students' oral presentation skills within the course structure, encompassing formats such as brief talks on current events in the data science and artificial intelligence field, literature discussions, ignite talks, and final project presentations. The inclusion of these elements allows for a comprehensive development of students' presentation abilities. Following multiple speaking engagements throughout the semester, the majority of students demonstrated an increased level of confidence and developed practical proficiency in professional public speaking.

Assessment To evaluate students' final reports, we have designed a comprehensive rubric, as presented in Table 2. For the literature presentation, we use a peer-evaluation form comprising four equally weighted dimensions: (1) breadth of topic coverage and organizational clarity, (2) depth of understanding as demonstrated in the presentation, (3) clarity of the delivery, and (4) presentation style, including multimedia usage, audience engagement, and overall interest level. For the final project presentation and demonstration, we resort to a similar peer-assessment methodology. It includes three equally weighted categories: (1) presentation, encompassing style, clarity, and accuracy, (2) creativity, measuring novelty in design, approach,

Table 2 Writing rubric for final term project report (total: 20 points)

	Needs improvement	Approaches expectations	Meets expectations	Excellent
Topic (4 pts.)	Poorly stated problem or unclear broader impact	Lacks clarity in problem or broader impact	Clear definition of problem and broader impact	Clear, well-articulated problem and broader impact. The value of the work is convincing
Outcomes (6 pts.)	Minimal description or lack of novelty in technical solution or result	Unclear description or inadequate justification of technical solution. Some effectiveness demonstrated. Lack of novelty	Clear description and justification of technical solution. Demonstrated effectiveness. Minimal novelty	Excellent discussion of technical solution and results. Good novelty. Exceptional insight into the solution's relevance to the problem. Strong evidence of effectiveness
Structure (6 pts.)	Fragmented and difficult to read. Does not provide a general picture of the project	Difficult to follow or verbose, repetitive descriptions	Easy to follow, clear descriptions, reasonably logical arguments	Clear, concise, easy to follow. Logical arguments. Appropriate section lengths
Grammar (2 pts.)	Unacceptable number of typos or grammatical errors	Noticeable typos or grammatical errors	Minimal typos or grammatical errors	Free from typos or grammatical errors
Citations (2 pts.)	Fails to use or incorrectly uses sources. Potential plagiarism will incur severe penalties	Uses sources but substitutes them for original ideas. Inconsistent citation format	Uses sources to support ideas. Appropriate quoting and paraphrasing but inconsistent citation format	Uses sources to support, not substitute for, original ideas. Skillful integration of published literature. Consistent citation format

or problem addressed, and (3) functionality/results, assessing the flawless operation of functions, the professional look and feel of the project, and the adequacy of the functions offered for the chosen data science application.

Problem-Solving Skills: Creative Thinking, Strategic Analysis

The ultimate objective of this course is to develop students' problem-solving capabilities. The complexity of real-world data science problems often lies in their multifaceted nature, characterized by the vastness of data, diversity in data types (including multimedia, sensor data, and graphs), imperfect data quality (such as noisy, incomplete, or inconsistent data), dynamic data attributes, and domain relevance. Additionally, data scientists have to continuously address various practical constraints or requirements, which may include appropriate modeling (for instance, avoiding overfitting or underfitting, parameter tuning), the interpretability/explainability of models, privacy and security considerations, ethical issues, and model maintenance tasks (like dynamically updating models as new data emerges).

Contrary to some other computing domains, such as numerical simulation and traditional databases, where a mathematical proof can confirm the correctness of an approach, the primary challenge in data science is the comprehensive orchestration of the entire pipeline. Therefore, identifying the optimal sequence of steps to constitute a valid pipeline is the focal point of this part of the course.

Creative Thinking Because of the complexity and nature of real-world data science problems, successful data scientists often need to devise novel solution pipelines. This underscores the value of pursuing data science as an undergraduate major, especially considering the abundance of open-source machine learning and data processing packages, coupled with the increasing sophistication of automated machine learning (AutoML) (He et al., 2021). We want to foster an environment where students can independently conceptualize and execute a complete pipeline, requiring minimal supervision from the course team.

This process starts with the *selection of project topics*. We encourage students to exercise their creativity by choosing their own project topics. While we are available to provide a "safety net" by suggesting topics, it is noteworthy that nearly all teams can formulate their own interesting ideas. We have communicated our expectations to them clearly: suitable project topics must integrate computational methods, mathematical/statistical approaches, and domain knowledge. Additionally, these topics should exhibit sufficient complexity and scale, have real-world implications, and feature some innovative aspect(s).

The following is a sample of twenty project topics conceived by our students:

- News topic analysis based on text data
- Modeling feature representations for human emotional states
- Color histogram matching approach for speech emotion recognition
- High-energy particle tracking in CERN detectors

- Decoding neuron interactions: predicting brain activity from semantic stimuli using fMRI
- Utilizing growth rings and PDSI to predict climatic events
- Natural language analysis of Twitter for the 2020 US Presidential Election
- Predicting Major League Baseball pitcher injuries with Statcast data
- Classification of maritime vessels using feature engineering
- Analysis of the intersection between mental health, nutrition, and happiness
- A deep learning approach for malware classification
- Sports moneyline odds prediction using machine learning
- Analysis of homeless youth data to predict drug usage
- Automated identification of Spotted Lanternflies (Lycorma delicatula)
- Computer vision-based yoga posture recognition and rating system
- Speech-to-speech translation with adaptable text to speech
- Multimodal sarcasm detection in sitcoms
- Using explainable AI to interpret complex breast cancer detection models
- Transient detection in "nearby" galaxies using TESS data-driven pipeline
- Detect early-onset Parkinson's disease

This array of project topics covers several interdisciplinary fields including natural language processing (NLP) and text analysis, emotional and psychological analytics, speech and audio processing, neuroscience and biomedical studies, physics and astronomy, climate science and dendrochronology, sports analytics, cybersecurity, maritime science, social studies, entomology, and health and fitness.

Throughout the term project, we underscore the importance of novelty and creativity. We adopt an *apprenticeship* approach in our coaching, rather than the conventional production pipeline model used in typical lecture-based courses. We mandate students to incorporate an innovation segment in their project and to argue for that in their project reports, thus motivating them to continually evaluate the novelty of their project. While we provide necessary guidance and relevant academic papers when they need them, the responsibility is largely on the students to brainstorm their own solutions as a team.

When presenting their project, we also encourage students to be creative. They need to devise strategies to engage their peers while ensuring the inclusion of all necessary content within a limited amount of time. Students have used animations, self-recorded videos, and real-life demos, among other techniques.

Strategic Analysis The field of data science is currently in a state of explosive development, with new methods and challenges constantly emerging, often superseding their predecessors. To keep pace with these ongoing advancements, having refined academic acumen and insight is important. Data scientists need to have the capability to identify impactful R&D topics, discern datasets with wide-ranging potential benefits, and recognize methods with enduring value.

Developing such an insightful perspective is not an overnight task. In our course, we facilitate the nurturing of this outlook by discussing the motivations behind some classic projects. We provide students with "seed" academic papers to help initiate

their literature reviews and projects, thereby intentionally training them in strategic analysis.

Furthermore, we incorporate an exemplary real-world industrial project in the discussion to expose students to the strategic analysis done by leading data scientists in the industry when they face challenging problems. Specifically, we let students watch a video talk by Dr. Dhiraj Joshi of IBM Research, summarizing his work related to emotion understanding from videos (ISTatPENNSTATE, 2018). Students are then tasked with summarizing three key takeaways from both the talk's technical content and its delivery. According to their feedback, they have learned strategic analysis skills and formed new perspectives on some classic learning techniques such as the support vector machine (SVM) (Cortes and Vapnik, 1995), once they understood their practical applications. Some students expressed doubts about the extent of information one could extract from a dataset, while others generated ideas on evaluating the success of an AI system. The students also gleaned professional presentation skills from the video, including the presenter's style, use of gestures, incorporation of video content, and smooth transitions between topics.

Decision-Making Efficient and rational decision-making constitutes a vital component in the personal and professional growth of data science students. Data science projects often necessitate extensive resource management, spanning time, financial investment, data, computational resources, and manpower. Upon performing strategic analysis during problem selection, data gathering, and data processing phases, it is necessary to make informed decisions that maximize positive outcomes and mitigate risks.

This skill is developed during the entirety of the term project. Students are tasked with not only determining their primary topics but also the strategies employed to overcome project challenges, leveraging the resources available to them. In a data science pipeline, a decision is involved in every step, and these decisions can have an impact on the quality and validity of the work. Furthermore, students are invited to pose questions during in-person project meetings with the instruction team, scheduled twice in the semester. We offer guidance to ensure their progress when they encounter hesitations or face problematic decisions. This approach progressively expands students' competency in effective decision-making.

Table 3 provides a summary of the contributions of individual course components to the development of each of the four major skills. While our course facilitates the development of diverse skills, there remains room for reflection and improvement. The subsequent section will address these aspects from both the student and instructor perspectives.

Reflections and Lessons Learned

Student Feedback A news article initially covered the course's introduction, featuring a few student comments as testimonials (Hallman, 2018). Among them

Table 3 Development of diverse skills through various components of the course

Component	Skills			
	Technical	Intrapersonal	Interpersonal	Problem-solving
Short presentation	Emerging technologies	Self-confidence, Personal motivation	Public speaking	Information collection
MPI, PyTorch, Cluster/Cloud computing	UNIX, programming			
Time series	Statistical method			
Writing lectures and Workshop			Technical writing	Research and Information collection
Industrial invited speaker talk	Industrial applications of technologies		Business communication	Problem solving, Critical thinking
Literature review	Domain knowledge, Machine learning	Self-learning, Self-confidence, Personal motivation	Scholarly communication	Research skills, Analytical thinking
Readiness assessment tests	Domain knowledge, Machine learning			Project pipeline, Analytical thinking
Ignite talk			Public speaking, Self-confidence	
Project proposal			Team collaboration, Proposal writing	Idea generation, Creativity
Project and Reports	Data analytics, Visualization, Domain knowledge	Self-reflection, Personal motivation, Time management	Team management, Leadership	Critical thinking, Research skills
Peer review activities			Feedback and Communication	Assessment and Evaluation

was a statement from Ryan Jaeger, who was a second-year student when he took the course. Jaeger articulated his understanding of the course's objectives, stating, "In my mind, a good data scientist is able to manipulate data and perform the machine learning modeling necessary to answer a given question. However, a great data scientist is also able to generate meaningful research questions, place their research in context, and clearly communicate the significance and novelty of their work to many audiences. I believe the emphasis on communication in DS 340W will help me to become a great data scientist."

We consistently gather student feedback and course ratings at the end of each semester to continuously refine the course experience. Overall, students have expressed satisfaction and enthusiasm about the course. For instance, in the fall semester of 2022 (following the COVID-19 pandemic, but prior to the introduction of ChatGPT), the authors taught the course and received positive ratings: the course achieved a median score of 6 out of 7, while the instructor received a full score of 7 out of 7. These scores were based on responses from 72% of the class. Written feedback from students further revealed various aspects of the course that positively influenced their learning experience:

- **Hands-on Experience:** Students valued the practical, applied nature of the course, which offered firsthand experience in data science.
- **Reviewing Literature:** Analyzing articles and other people's methods helped students understand current industry practices and facilitated the application of these methods in other academic settings.
- **Freedom and Creativity:** Students appreciated the liberty to choose their own topics for open-ended projects, which made the learning process more enjoyable and enriching.
- **Focus on Presentation Skills:** The course not only developed data science skills but also nurtured presentation abilities. Despite initial fear, students recognized the importance of public speaking and appreciated the focus on presentation skills.
- **Writing Workshops:** Workshops and one-on-one feedback from the professor and TA were deemed very helpful. Students learned academic writing techniques specific to data science.
- **Use of Tools:** Tools like Overleaf for writing formal reports were beneficial for students' learning.
- **Literature Talk:** Activities like literature talk were appreciated for boosting confidence in public speaking.
- **Professor's Enthusiasm:** The professor's passion for the subject matter was noted as a key factor that made the class more engaging and information more retainable.
- **Class Attendance and Interesting Talks:** Regular attendance and exposure to intriguing talks made the learning process more effective and enjoyable.

The students suggested improvements for enhancing their learning experience, including the following:

- **RATs:** Some students suggested that RATs should be replaced with class discussions, where every student has an opportunity to participate. Some students found the grading standards too stringent.
- **LaTeX:** While acknowledging the usefulness of the Overleaf tool, several students found it challenging to use. They proposed holding workshops or dedicated classes to teach students how to use Overleaf effectively.

- **Project:** Some students suggested replacing the single, semester-long project with multiple smaller projects. This would offer more opportunities for presentations and writing, thus enhancing their learning.
- **Programming:** Students asked for more programming assignments to help them learn useful technologies. Some students found the supercomputer assignments challenging and suggested more demonstrations and hands-on practice.
- **Writing:** A few students found the feedback on papers and presentations to be vague and requested more detailed and constructive feedback.

Successes and Challenges From the instructors' perspective, the course has achieved its goals in several key areas. It has successfully instilled in students the significance of data science, guided them through the typical project pipeline, enabled them to comprehend the challenges involved in implementing each pipeline step correctly, and cultivated their problem-solving skills for handling large and complex real-world problems. It has also developed their communication abilities. We are pleased that most students can complete a project using a robust methodology and articulate their findings both in writing and orally.

The course's effectiveness and impact are further substantiated by the feedback we've received from past alumni of the program. They've noted that this is one of the few courses where the acquired skills can be directly applied in a professional setting as a data scientist.

However, in spite of our continuous efforts to enhance the course, certain challenges persist such as the following:

- **Balancing Topics:** It is difficult to cover so many different topics in one semester. We dedicate a substantial portion of class time to student presentations to provide them with opportunities to present formally and informally in front of a live audience. This inevitably reduces the time for lectures or hands-on programming, which is reflected in the subpar performance of some students in programming assignments.
- **Teamwork Issues:** Some students, overwhelmed with multiple courses, struggle to contribute equally to team projects. As a result, we often encounter issues such as uneven contributions or students only taking part in tasks they are already proficient in.
- **Plagiarism:** We use Turnitin to detect plagiarism in final reports. Every semester, despite we emphasize academic integrity repeatedly, we can detect plagiarizing. With AI-based writing tools like ChatGPT and paraphrasing tools becoming more available, plagiarism detection becomes even more challenging. Our current approach allows students to use these tools to enhance readability and language quality, but they must disclose such use. Furthermore, using ideas from published work without appropriate attribution or citation is deemed plagiarism. However, it's important to note that detecting this type of plagiarism can be extremely difficult.
- **Artistic Sense:** The development of esthetically pleasing visualization and talk slides is a crucial aspect of communication skills in data science. However,

nurturing an artistic sense in students within a short semester proves to be a daunting task. This is reflected in their reports and presentation slides. We have encouraged them to learn from professionals by reading high-quality published articles and adopting professional slide templates, but observable improvements have been minimal over the semester. We believe data science students should be encouraged to engage more with artworks, enroll in art classes, practice creating art, and/or analyze artworks to gradually develop their artistic sense.

- **Risk-Taking and Novelty:** We encourage students to embrace risk-taking and innovation when selecting their project topics. However, our observation is that most teams tend to be more conservative. This cautious approach might be attributed to this being their first experience with real-world problems. It may require multiple such exposures before they can gain more confidence and venture into unexplored territories.
- **Insufficient Scale or Complexity:** Although we provide students with access to supercomputers and occasionally commercial cloud platforms, only a handful of project teams managed to work with data of a scale and complexity that justified the use of such resources. Consequently, many students did not significantly enhance their skills in managing large-scale, complex problems through this course.
- **Fair Peer Reviewing:** Generally, students are fair and constructive in evaluating their classmates' presentations. However, we have occasionally observed variations in scoring where some students consistently provide low scores, while others give exceptionally high scores. We attribute these discrepancies not to any malicious intent but rather to the fact that different individuals have varying criteria for assessing others' work.

Concluding Remarks

The process of teaching this course has been a continual learning experience. It has refined our understanding of effective pedagogy for teaching data science, guided by evolving trends in the field, emerging educational methodologies, and most importantly, the feedback and progress of the students themselves. Designing and implementing a project-focused, writing-intensive data science course that nurtures an impactful and engaging learning experience requires a careful balance of theoretical knowledge and practical application, individual growth and collaborative engagement, and adhering to established norms and promoting innovation. Seeing our students undertake intensive explorations of complex concepts, apply what they've learned to real-world challenges, and steadily grow into emergent data scientists has been an enormously gratifying affirmation of our teaching methodology's effectiveness.

Acknowledgments We extend our appreciation to Jia Li, who codeveloped and co-instructed the course in its formative years. We also thank Tongan Cai, Yukun Chen, Dolzodmaa Davaasuren, Yu Li, Benjamin Wortman, and Sitao Zhang, who have contributed to the course development and/or delivery in their capacity as teaching assistants. The course used the Extreme Science and Engineering Discovery Environment, which was supported by the National Science Foundation (NSF) under Grant No. ACI-1548562, the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) Program supported by NSF under Grant Nos. OAC-2138259, OAC-2138286, OAC-2138307, OAC-2137603, and OAC-2138296, and the Amazon Web Services through the Educate Program. The data science research of the authors is supported by the NSF under Grant No. 2216127. The data science research of Wang is also supported by the NSF under Grant Nos. 2234195, 2205004, and 2015943, the National Institutes of Health under Grant No. R01EB030130, and the National Endowment for the Humanities under Grant No. HAA-287938-22. Wang would also like to acknowledge the Amazon Research Awards program for the generous gifts (2019-2023).

References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European court of human rights: A natural language processing perspective. *PeerJ Computer Science, 2*, e93.

Business Information Review Editorial (2018). The importance of continuous learning for innovation, progression and survival. *Business Information Review, 35*(1), 6–8.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., & Leonard, J. J. (2016). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics, 32*(6), 1309–1332.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172–186.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning* (pp. 1597–1607). PMLR.

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 2094–2107.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods Natural Language Processing* (pp. 1724–1734).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision, Part III* (pp. 288–301). Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning* (pp. 1126–1135). PMLR.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423).

Hallman, J. (2018). Data sciences program blends interdisciplinary training for a growing industry. *Penn State News*.

He, X., Zhao, K., & Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems, 212*, 106622.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *Proceedings of NIPS Deep Learning and Representation Learning Workshop*.

ISTatPENNSTATE (2018). *Penn State Startup Week 2018 - Dr. Dhiraj Joshi, Senior Research Scientist, IBM Watson*. https://www.youtube.com/watch?v=j5r0vaXQCr8

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*.

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *Proceedings of the Deep Learning Workshop, in Conjunction with the International Conference on Machine Learning* (Vol. 2). Lille.

Li, H., Ota, K., & Dong, M. (2018). Learning IoT in edge: Deep learning for the internet of things with edge computing. *IEEE Network, 32*(1), 96–101.

Li, J., Yao, L., Hendriks, E., & Wang, J. Z. (2012). Rhythmic brushstrokes distinguish van Gogh from his contemporaries: Findings via automated brushstroke extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(6), 1159–1176.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99–106.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2574–2582).

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks, 22*(2), 199–210.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems, 28*.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285–295).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems, 27*.

U.S. Bureau of Labor Statistics (2023). *Data Scientists*. https://www.bls.gov/ooh/math/data-scientists.htm. [Online; accessed 14-June-2023].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*.

Weiser, O., Kalman, Y. M., Kent, C., & Ravid, G. (2022). 65 competencies: Which ones should your data analytics experts have? *Communications of the ACM, 65*(3), 58–66.

Zhang, Z., Li, J., Stork, D. G., Mansfield, E., Russell, J., Adams, C., & Wang, J. Z. (2022). Reducing bias in ai-based analysis of visual artworks. *IEEE BITS the Information Theory Magazine, 2*(1), 36–48.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2223–2232).