10-3-2019

Integrating Data Science into a General Education Information Technology Course: An Approach to Developing Data Savvy Undergraduates

Malcolm Haynes
United States Military Academy, malcolm.haynes@westpoint.edu

Joshua Groen
United States Military Academy

Eric Sturzinger
United States Military Academy

Danny Zhu
United States Military Academy

Justin Shafer
United States Military Academy

See next page for additional authors

Recommended Citation

Malcolm Haynes, Joshua Groen, Eric Sturzinger, Danny Zhu, Justin Shafer, and Timothy McGee. 2019. Integrating Data Science into a General Education Information Technology Course: An Approach to Developing Data Savvy Undergraduates. In Proceedings of the 20th Annual SIG Conference on Information Technology Education (SIGITE '19). Association for Computing Machinery, New York, NY, USA, 183–188. DOI:https://doi.org/10.1145/3349266.3351417

Authors

Malcolm Haynes, Joshua Groen, Eric Sturzinger, Danny Zhu, Justin Shafer, and Timothy McGee

Integrating Data Science into a General Education Information Technology Course

An Approach to Developing Data Savvy Undergraduates

Malcolm Haynes
United States Military Academy
West Point, New York
malcolm.haynes@westpoint.edu

Joshua Groen
United States Military Academy
West Point, New York
joshua.groen@westpoint.edu

Eric Sturzinger
United States Military Academy
West Point, New York
eric.sturzinger@westpoint.edu

Danny Zhu
United States Military Academy
West Point, New York
danny.zhu@westpoint.edu

Justin Shafer
United States Military Academy
West Point, New York
justin.shafer@westpoint.edu

Timothy McGee
United States Military Academy
West Point, New York
timothy.mcgee@westpoint.edu

ABSTRACT

The National Academies recommend academic institutions foster a basic understanding of data science in all undergraduates. However, data science education is not currently a graduation requirement at most colleges and universities. As a result, many graduates lack even basic knowledge of data science. To address the shortfall, academic institutions should incorporate introductory data science into general education courses.

A general education IT course provides a unique opportunity to integrate data science education. Modules covering databases, spreadsheets, and presentation software, already present in many survey IT courses, teach concepts and skills needed for data science. As a result, a survey IT course can provide comprehensive introductory data science education by adding a data science module focused on modeling and evaluation, two key steps in the data science process. The module should use data science software for application, avoiding the complexities of programming and advanced math, while enabling an emphasis on conceptual understanding. We implemented a course built around these ideas and found that the course helps develop data savvy in students.

CCS CONCEPTS

• **Information systems → Data mining**; • **Social and professional topics → Information systems education**; **Information technology education**;

KEYWORDS

Introductory Data Science, Information Technology Education, Database, Data Mining, Data Visualization, CRISP-DM

ACM Reference Format:
Malcolm Haynes, Joshua Groen, Eric Sturzinger, Danny Zhu, Justin Shafer, and Timothy McGee. 2019. Integrating Data Science into a General Education Information Technology Course: An Approach to Developing Data Savvy Undergraduates. In *The 20th Annual Conference on Information Technology Education (SIGITE ’19), October 3–5, 2019, Tacoma, WA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3349266.3351417

1 INTRODUCTION

Every day 2.5 quintillion bytes of data are created [12]. The explosive growth of data has created new opportunities for increased productivity across many industries. Leveraging the opportunities requires data science, the use of scientific methods, processes, algorithms and systems to extract knowledge and insights from data [11]. Personnel who have a basic understanding of data science are needed globally across a wide range of organizations [30]. In the US alone, McKinsey & Co. projected a need for 1.5 million additional data savvy workers “who can ask the right questions and consume the results of the analysis of big data effectively” [28]. The need for a data savvy workforce suggests that data science education should be integrated into the core undergraduate curriculum. However, there is little appetite to add a core course in data science at most institutions. A potential alternative is to integrate a module on data science within an existing general education course. Many institutions already offer an introductory information technology (IT) course as part of the general education curriculum. A general IT course is uniquely suited to integrate a data science module since it may already include lessons on databases, spreadsheets, and presentation software-topic areas that are often required to practice data science.

We present an approach to integrating a data science module within a general education IT course. Our approach is based on three key ideas. First, organize learning around a data science process such as CRISP-DM. Next, within each step of the process, focus learning on key data science concepts. Finally, have students execute the process using GUI-based IT tools. Such an approach is appropriate for a module within a survey course and helps obviate a requirement for students to have deep statistical knowledge and programming experience.

In academic year 2019, we integrated data science into our general education IT course as part of a comprehensive survey of practical computing knowledge. We evaluated the course with both quantitative and qualitative methods. Results indicate that

SIGITE ’19, October 3–5, 2019, Tacoma, WA, USA
2019. ACM ISBN 978-1-4503-6921-3/19/10.
https://doi.org/10.1145/3349266.3351417

our course fosters data savvy. In addition, students express the class material has utility outside of class.

The rest of the paper is organized as follows: Section 2 discusses background material relevant to data science education, Section 3 describes the principles on which we base the course and course components that satisfy those principles, Section 4 explains how we evaluated and assessed our course, and Section 5 concludes the study.

2 BACKGROUND

The National Academies of Sciences, Engineering, and Medicine (NASEM) recommend that academic institutions "encourage the development of a basic understanding of data science in all undergraduates" [33]. The NASEM call presents challenges for academic organizations because 1) data science is a difficult topic to teach and 2) institutions would have to modify their general education curriculum to reach all students.

2.1 Introductory Data Science Challenges

Data science can be a challenging subject to teach. Two required data science skills, computer science (specifically, programming) and statistics can be difficult subjects for many students [15, 16, 37, 39]. A key takeaway from the literature is that the most common grade in an introductory programming course is an "F" [38]. This may be because concepts must be connected in increasingly complex ways, making it difficult to succeed if early ideas are not fully understood. Likewise, 34% of students taking an introductory statistics class either drop out or fail the course [13]. Negative attitudes and a lack of interest, among other reasons, help explain the low success rate [41].

In addition to subject matter difficulty, the current general education curriculum at most schools does not expose all students to data science. Schools that offer introductory data science courses often allow them to satisfy a general math or statistics requirement. However, a study of 50 colleges and universities found that 62% do not require students to take mathematics [27]. Of those that do require math, classes not related to statistics (e.g. college algebra, trigonometry, calculus) can often be used to satisfy the requirements [4]. In a similar vane, data science can be used to satisfy a science and technology requirement at some institutions. However, 38% of schools do not require a science/technology course as part of the general education requirements [27].

Finally, deciding what to teach in an introductory data science class can be difficult. There are few published standards or learning objectives for undergraduate data science courses for non-data science majors [24]. The undergraduate curricula that do exist are generally aimed at data science majors [1, 8]. Recent work is attempting to generate useful guidelines [7, 8, 10, 34], but data science pedagogy for the general student populace is still in its nascent stages.

2.2 Introductory Data Science Courses

Notwithstanding the challenges of incorporating data science, a growing number of institutions are starting to offer introductory data science education via "a single inspirational course that could satisfy a general education requirement" [33]. Most introductory data science courses tend to have either a statistics or a computation focus.

Much of the literature on introductory data science education focuses on statistical reasoning [14, 17, 18, 20, 21, 31, 33, 43]. A representative example of a statistics focused course is "Reasoning with Data" offered at Carnegie Mellon University (CMU). The course focuses on making statistical decisions using data. The majority of the course covers traditional statistics topics such as probability, sampling, and hypothesis testing. It can be taken in lieu of a traditional statistics class to satisfy a general education requirement for students in CMU's college of humanities and social sciences [34, 35].

Another approach to introductory data science courses focuses on computation [2, 5, 33]. An example of this approach is "Introduction to Data Science" offered by the University of Illinois at Urbana-Champaign (UIUC). The course uses the Python programming language and covers a wide range of topics including, but not limited to, general programming concepts, data preparation, relational databases, machine learning, and data visualization [5]. There are no prerequisites, but the course authors recognize the difficulties of teaching programming within a data science course, noting that the heavy programming requirements "could turn away students who are new to programming" [5]. Authors of similar courses report similar difficulties [2].

Of note, we found no examples in the literature of a data science module integrated into an introductory IT course.

2.3 Opportunities

Because most institutions have numerous courses that can satisfy general education requirements, data science should be integrated into more general education courses. Almost all the cited authors of introductory data science courses recognize that specific curriculum development is in its infancy and describe their courses as a snapshot of an evolving product. However, using statistics or computation focused data science courses to satisfy general education requirements is necessary but not sufficient to ensure all graduates are exposed to data science. Although many introductory data science courses have no prerequisites, the extensive use of coding or statistics may dissuade some students from taking the course when less technical alternatives are available. Therefore, students need a data science option that does not rely on advanced math or programming.

There is no consensus on what all graduates should know about data science. Terms such as "basic understanding", "data acumen", and "data savvy" are present in the literature, often without clear definitions. A survey of the term "data savvy" results in key attributes listed by multiple authors. The most commonly listed attribute is the ability to ask the right questions - questions that, if answered, will help solve an organization's problems [3, 25, 26, 29]. Asking the right questions does not require programming or advanced math [29]. However, it does require some practical knowledge of the data science process [25, 26]. Thus, we define "data savvy" as a practical knowledge of the data science process and comfort using its results to make data-driven decisions. An entire course on data science may not be required to foster data savvy; a

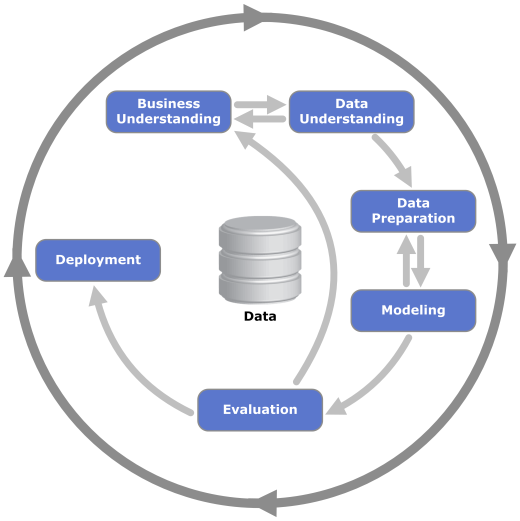

Figure 1: CRISP-DM Process Diagram [23]

well-constructed data science module within a broader course may be enough.

Many introductory IT courses constitute a survey of practical computing knowledge [40, 42] and include modules on spreadsheets, databases, and presentation software - topics which provide a foundation for doing data science in practice. A module on data science in a general education survey IT course can build upon standard IT topics. Further, data science software can enable application without requiring advanced math or programming and permit the course to focus on key concepts. Integrating data science into a survey IT course provides a way to expose more students to data science and helps develop data savvy.

3 INTEGRATING DATA SCIENCE

A survey IT course is uniquely situated to integrate data science and can help students become more data savvy. In particular, an IT course can introduce students to a data science process that students can use to solve problems. The cross-industry process for data mining (CRISP-DM) is the most commonly used data science process [36] and consists of six iterative, non-linear phases (Figure 1). There are entire courses devoted to each phase. However, a survey IT course can help students gain a general understanding of the data science process. Databases, a traditional IT topic, are often key components in two of the CRISP-DM phases - data understanding and data preparation - helping students quickly grasp those phases. Further, data science software permits abstraction of complex implementation details for all phases, allowing students to focus on conceptual understanding while still being able to implement the process from beginning to end.

In academic year 2019, we revised our general education survey information technology course "Cyber Foundations" by integrating data science. We created a six lesson module organized around the CRISP-DM methodology, applied using RapidMiner, a GUI-based data science platform. Each lesson in the module lasts 75 minutes and is generally split into two components. The first component consists of teacher-led instruction. The purpose is to give students an intuitive explanation of complex concepts. In the second portion of the class, the concepts discussed previously are applied via worked examples with forward-fading [32]. After each lesson, students are assigned homework problems designed to take the average student about an hour to complete. In addition to worked-examples, we utilize project-based learning and assign a group project that requires the practical application of the data science process to complete. To create space for the data science module, we eliminated material on digitization and reduced coverage of networking.

3.1 Supporting Modules

Many survey IT courses already provide a foundation for data science. In particular, database modules teach concepts associated with data understanding and data preparation. As a result, students are able to apply previously learned concepts and quickly grasp the data understanding and data preparation phases of CRISP-DM.

Data preparation includes, in part, selecting, merging, deriving, reformatting, and summarizing data [6], all traditional database concepts. The core of data understanding is exploring data using queries (a standard database practice) and visualization to understand relationships between attributes, view results of simple aggregations, and describe the data with simple statistical analyses [6].

Our course includes a database module. Our database module includes lessons on database design and queries. We minimize the need for programming during the database module through the use of Microsoft Access, a graphical database management system. We chose to use Access because it is does not require learning SQL and it is widely used throughout industry, being the only GUI-based platform with more than 10% of the total market share for database management systems [22]. Access allows users to build a relational database and create queries through a combination of drag and drop interactions, drop down menus, and functional statements.

Data exploration also uses visualization. While data visualization blocks are not a common element in survey IT courses, many courses have modules on spreadsheets and presentation software. Both types of software often include the ability to create simple data visualizations. Our course includes a block on data visualization. While the block primarily focuses on communicating data, students are taught techniques appropriate for exploratory visualization such as scatterplots with trendlines.

3.2 Data Science Module

A way to increase data savvy is to educate students on the data science process, provide an intuition on key concepts within the process, and have students apply their knowledge using problems and projects.

Of the six phases of the CRSIP-DM process (Figure 1), an overview alone is perhaps most appropriate for business understanding since many of the details of the phase - requirements, assumptions, constraints, costs, benefits, and more - either require specific domain knowledge or are more suitable for discussion in a business class. In addition, an overview may be appropriate for data understanding if students have been exposed to databases and data visualization in

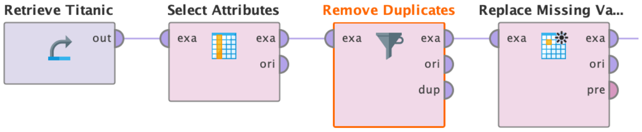

Figure 2: Sample RapidMiner data cleaning process.

other parts of the course. Our course provides an overview only of the business understanding and data understanding phases since students complete modules on databases and data visualization as part of the course.

3.2.1 Data Preparation. While a database module may cover much of data preparation, it is unlikely to sufficiently address data cleaning - an essential task for data preparation. Therefore, a data science module will likely need to address data cleaning in some detail.

In our course, the data science module starts with an introduction to data science and the CRISP-DM methodology. After the overview of CRISP-DM, we focus on data preparation. We discuss many of the key tasks of data preparation and link them back to concepts and skills learned in the database module. Next, the cleaning step of data preparation is addressed in detail and culminates with worked examples of data cleaning using RapdMiner. An example data cleaning process is shown in Figure 2.

RapidMiner is a data science platform that uses a GUI to "empower non-experts to get the same findings as data scientists" [19]. We use RapidMiner because it offers a free educational license and does not require any programming knowledge. In addition, the RapidMiner GUI is intuitive and easy to use. In RapidMiner, each operation in a data science process is represented by a block that students can drag-and-drop into the work environment as they build their solution. Parameters of each functional block are set by a text box or drop-down menu.

3.2.2 Modeling. The modeling phase is perhaps the most mathematically demanding in the CRISP-DM process. Models are based on the application of advanced statistics that can be daunting for some to understand. Therefore, when introducing modeling, the focus should be on understanding how different models work at a conceptual level. Students should then employ software to build models followed by using the built models to classify or regress unlabeled data. We believe building and using models is a key aspect in making students more comfortable using the results of data science to make data-driven decisions.

In our course, the modeling instruction begins with introducing students to the concepts of supervised and unsupervised machine learning. We also introduce students to the concepts of regression and classification, noting potential business use cases for both. In the module, we focus on classification methods, specifically decision trees, logistic regression, and k-nearest neighbors. For a given technique, we first focus on how it works at an intuitive level and then provide insight into the underlying mathematics. After developing a base understanding for a machine learning technique, students build models based on the technique using RapidMiner.

The modeling phase of the CRISP-DM methodology includes model assessment. We start teaching model assessment by discussing overfitting and the trade-offs between model complexity and generalization. Next, we introduce students to the ideas of testing and holdout datasets as well as cross validation. We then discuss different types of errors and associated costs which leads into the expected value framework for assessing models. Finally, students are introduced to the Receiver Operating Characteristic (ROC) curve and Area Under the ROC curve (AUC).

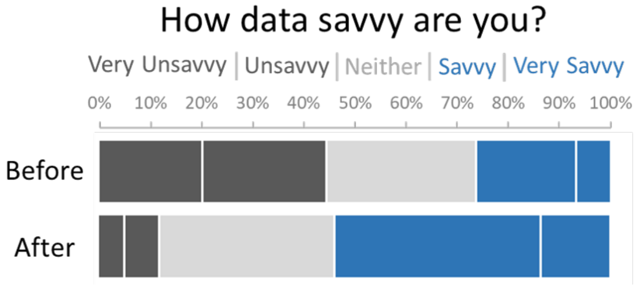

Figure 3: Change in student-reported data savvy (n=271). Based on averaging of Q2 and Q3

3.2.3 Evaluation and Deployment. Understanding the first four phases of the CRISP-DM process are essential for becoming more data savvy. Therefore, an overview of the final two phases is sufficient to complete an introduction to data science. The core of the evaluation phase is to determine if the model solves the stated business problem while the deployment phase puts the model into a production environment [6]. Of note, part of the deployment phase is communicating the results of the entire CRISP-DM process. Our course includes a module on data visualization focused on communicating data to leaders and decision makers.

3.2.4 Project. A comprehensive project is a way to reinforce and find gaps in learning. In our course, students are assigned a group project. The project requires applying the CRISP-DM process using data science software. In the spring semester of 2019, students were provided individual player statistics for every season of the NBA through 2016. The data included missing and corrupted examples. The goal was to predict, based solely on rookie year statistics, which athletes would play for five or more years. To complete the project, students had to clean the data set, compare and contrast the classification performance of three different models, and recommend a model for deployment. The final deliverable was a presentation detailing the group's approach to each step of the data mining process, to include discussion of model assessment, and rationale for the final recommendation.

4 COURSE EVALUATION AND ANALYSIS

Data savvy requires practical knowledge of, and experience with, the data science process and associated tools. We used a survey to gauge student data savvy. In addition, we examined performance on projects and exam questions to evaluate practical knowledge.

4.1 Survey

In the 2019 spring semester, students were surveyed at the start and end of the course to gauge their knowledge of, exposure to, and comfort with a variety of data science topics and applications. 392 student took the course, the majority (73%) of whom major in a

Quest.	μ_{pre}	μ_{post}	t-test		MWU-test	
			t	p-value	U	p-value
Q1	2.55	3.42	12.014	<0.001	21,164	<0.001
Q2	2.82	3.59	9.282	<0.001	23,392	<0.001

Table 1: Results of pre- and post- course survey comparison (n = 271) using paired t-test and MWU-test.

subject outside of STEM. 380 students answered the pre-survey, 287 answered the post-survey, and 271 students answered both. Survey questions used a five-point Likert scale, with five indicating students felt very confident or very comfortable with a particular topic. In the post-survey, students also had the opportunity to discuss if and how they used course material outside of class. Of the 21 survey questions administered before and after the course, two questions gauged students’ attitude towards data science:

Q1: *How confident are you that you can use data mining to perform predictive analysis?*

Q2: *How comfortable are you using the results of data mining to help make data driven decisions?*

A comparison of students’ self-reported comfort levels are summarized in Figures 3. The values in Figure 3 are based on averages of the responses to Q1 and Q2 above.

The pre- and post- survey attitude-related questions were analyzed using the t-test which is commonly used on Likert scale data (Table 1). Results indicate that students were significantly more comfortable with data science topics after taking the course than before. Data was also analyzed using the Mann-Whitney U (MWU) test with continuity correction which is commonly used to assess ordinal data [9]. The MWU test results confirm the t-test results.

In addition to attitudes, students were asked which kind of tool they would use to predict trends or classify data. Over 95% of students selected a spreadsheet before the course while less than 10% identified RapidMiner or similar software as an appropriate tool for predictive analysis and classification. After the course, students were able to select more appropriate IT tools for the question and over 65% of students indicated they would use a data mining tool like Rapid Miner for predictive analysis and classification tasks.

Students were also asked to discuss if and how they used course material outside of class. 46% of students indicated they used some course material outside of class. However, the majority of students did not use data mining. Comments from one student summarized the general sentiment-“too complicated for the small tasks we have to do outside of class.” Those who did report using data mining mostly used it in other academic classes. One student stated, “I have been using Rapid Miner on my thesis to try and inform my analysis in a way that I hadn’t considered looking at.” Other areas students reported using data mining were imagery analysis, data analytics, decision analysis, statistics, and various science classes. Despite not using data mining, 87% reported an increased or greatly increased ability to perform data mining.

4.2 Graded Results

Assignments associated with the data science module include five homework assignments, a quiz, an exam, and a project. In addition, material from the data science block comprised approximately 25% of the final exam. Exams consist of two principle components: conceptual questions and practicum questions. Conceptual questions are defined as those which test general understanding of key ideas. The practicum questions require students to use IT software in the execution of the data science process to solve a problem.

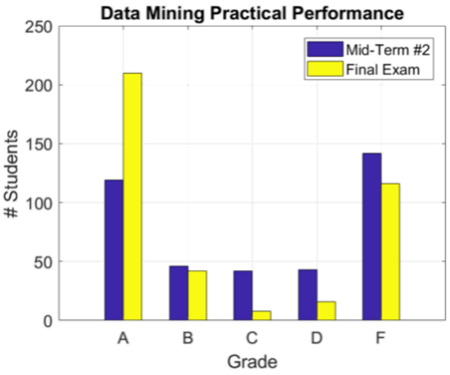

Figure 4: Comparison of Student performance on data mining practical during Mid-Term #2 and Final Exam (n=383).

An analysis of the exams indicates a gap between conceptual understanding and an ability to apply the concepts using IT. Students averaged 83% and 87% on data mining core questions on the midterm and final respectively, suggesting conceptual understanding and retention. However, average scores on practical questions were 70% and 76% on the midterm and final respectively, suggesting relative difficulty applying concepts.

In addition, one reason we chose to use a GUI-based data science tool was to avoid some of the challenges inherent in teaching programming. In an introductory programming course, the initial grade is the best indicator of the final grade. In our course, the results of the midterm practical resembled results for an introductory programming course [38]; the most common grade was an F followed by an A (Figure 4). However, unlike introductory programming courses, the initial grade was not the best indicator of performance on the final. We observed a noticeable shift in grade distribution. The number of students who earned an A on the final practical nearly doubled, suggesting the use of GUI-based tools helped students recover from initial poor understanding.

5 CONCLUSION

Data science is a difficult subject to teach, especially to a non-technical student cohort. Most data science courses require knowledge of computer programming and statistics, two subjects that previous research has shown are hard to learn. However, it is possible to foster data savvy within the general student populace by focusing learning on key data science concepts and abstracting programming syntax and statistical methods with GUI-based applications.

Teaching a limited set of critical concepts allows students to focus their cognitive effort. It also integrates well with a worked examples approach which lowers barriers for the acquisition of otherwise complex technical content. Furthermore, integrating data science into a survey IT course eased learning by reusing previously learned concepts.

In addition to understanding concepts, data savvy requires practical application which is normally done via programming. However, given the difficulties in learning programming languages, an introductory data science course should consider using GUI-based tools to apply concepts. As one student stated, "I didn't have to memorize code; instead I was visualizing pictures. It was easier for me to conceptualize what was happening to my data."

Using the described approach, we implemented an introductory data science course that helps foster data savvy. Students learned the CRISP-DM data science process and applied it on individual problems and a comprehensive group project. Results from student surveys and graded events suggest that our approach was successful in giving students a practical knowledge of data science and making them more comfortable making data-driven decisions.

ACKNOWLEDGEMENTS

Special thanks to Dr. Edward Sobiesk and Dr. Suzanne Matthews for their assistance in shaping and reviewing this paper. Their efforts significantly improved the final version. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Army, Department of Defense or the U.S. Government.

REFERENCES

[1] Paul Anderson, James Bowring, Renée McCauley, George Pothering, and Christopher Starr. 2014. An undergraduate degree in data science: curriculum and a decade of implementation experience. In *Proceedings of the 45th ACM technical symposium on Computer science education*. ACM, 145–150.

[2] Ben Baumer. 2015. A data science course for undergraduates: Thinking with data. *The American Statistician* 69, 4 (2015), 334–342.

[3] Christian Bonilla. 2015. Not a Data Scientist? You Can Still Be Data Savvy. http://www.smartlikehow.com/blog-native/2015/12/1/not-a-data-scientist-no-problem-you-can-still-be-data-savvy-than-most

[4] Brian Bourke, Nathaniel J Bray, and C Christopher Horton. 2009. Approaches to the core curriculum: An exploratory analysis of top liberal arts and doctoral-granting institutions. *The Journal of General Education* 58, 4 (2009), 219–240.

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

[6] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. 2000. CRISP-DM 1.0 Step-by-step. *ASHA presentation* (2000), 73. https://doi.org/10.1109/ICETET.2008.239 arXiv:arXiv:1011.1669v3

[7] Andrea Danyluk, Paul Leidig, Lillian Cassel, and Christian Servin. 2019. ACM Task Force on Data Science Education: Draft Report and Opportunity for Feedback. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, 496–497.

[8] Richard D De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S Baumer, Andrew Bray, Thomas C Bressoud, Lance Bryant, Lei Z Cheng, Amanda Francis, Robert Gould, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application* 4 (2017), 15–30.

[9] Joost CF De Winter and Dimitra Dodou. 2010. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation* 15, 11 (2010), 1–12.

[10] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, and Steve Brewer. 2016. EDISON data science framework: a foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 620–626.

[11] Vasant Dhar. 2013. Data Science and Prediction. *Commun. ACM* 56, 12 (Dec. 2013), 64–73. https://doi.org/10.1145/2500499

[12] Ciprian Dobre and Fatos Xhafa. 2014. Intelligent services for big data science. *Future Generation Computer Systems* 37 (2014), 267–281.

[13] Lawrence B Feinberg and Silas Halperin. 1978. Affective and cognitive correlates of course performance in introductory statistics. *The Journal of Experimental Education* 46, 4 (1978), 11–18.

[14] William Finzer. 2013. The data science education dilemma. *Technology Innovations in Statistics Education* 7, 2 (2013).

[15] Joan Garfield. 1995. How students learn statistics. *International Statistical Review/Revue Internationale de Statistique* (1995), 25–34.

[16] Joan Garfield and Dani Ben-Zvi. 2007. How students learn statistics revisited: A current review of research on teaching and learning statistics. *International statistical review* 75, 3 (2007), 372–396.

[17] Johanna Hardin, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, Roger Peng, Paul Roback, D Temple Lang, et al. 2015. Data science in statistics curricula: Preparing students to âĂIJthink with dataâĂİ. *The American Statistician* 69, 4 (2015), 343–353.

[18] Stephanie C Hicks and Rafael A Irizarry. 2018. A guide to teaching data science. *The American Statistician* 72, 4 (2018), 382–391.

[19] Markus Hofmann and Ralf Klinkenberg. 2013. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

[20] Nicholas J Horton, Benjamin S Baumer, and Hadley Wickham. 2014. Teaching precursors to data science in introductory and second courses in statistics. *arXiv preprint arXiv:1401.3269* (2014).

[21] Nicholas J Horton, Benjamin S Baumer, and Hadley Wickham. 2015. Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *arXiv preprint arXiv:1502.00318* (2015).

[22] IDatalabs. [n. d.]. Database Management System Products. https://idatalabs.com/tech/database-management-system/

[23] Kenneth Jensen. 2012. CRISP-DM Process Diagram. https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

[24] Jeremiah W Johnson. 2018. Scaling up: introducing undergraduates to data science early in their college careers. *Journal of Computing Sciences in Colleges* 33, 6 (2018), 76–85.

[25] Richard Kouri and Chad Morris. 2016. What Makes A Data-Savvy Manager. https://cims.ncsu.edu/cims_newsletter/mayjune-2016/what-makes-a-data-savvy-manager/

[26] Rho Lall. [n. d.]. Make Everyone Data Savvy, Forget Data Science. http://www.assume-wisely.com/data-savvy-managers-6-skills-tech-startups-look/

[27] Barry Latzer. 2004. The Hollow Core: Failure of the General Education Curriculum. A Fifty College Study. *American Council of Trustees and Alumni* (2004).

[28] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. (2011).

[29] Bernard Marr. 2015. Forget Data Scientists - Make Everyone Data Savvy. https://www.datasciencecentral.com/profiles/blogs/forget-data-scientists-make-everyone-data-savvy

[30] Betty Mbwilo, Honest Kimaro, and Godfrey Justo. 2019. Data Science Postgraduate Education at University of Dar es Salaam in Tanzania: Current Demands and Opportunities. In *International Conference on Social Implications of Computers in Developing Countries*. Springer, 349–360.

[31] Amelia McNamara, Nicholas J Horton, and Benjamin S Baumer. 2017. Greater data science at baccalaureate institutions. *Journal of Computational and Graphical Statistics* 26, 4 (2017), 781–783.

[32] Roxana Moreno and Richard E. Mayer. 2003. Cognitive Load Theory and Instructional Design: Recent Deve. *Educational Psychologist* 38, 1 (2003), 43–52. https://doi.org/10.1207/S15326985EP3801

[33] Engineering National Academies of Sciences, Medicine, et al. 2018. *Data science for undergraduates: opportunities and options*. National Academies Press.

[34] Engineering National Academies of Sciences, Medicine, et al. 2018. *Envisioning the data science discipline: the undergraduate perspective: interim report*. National Academies Press.

[35] Rebecca Nugent Philipp Burckhardt, Francis R. Kovacs and Ron Yurko. [n. d.]. ISLE: A Browser-Based E-Learning Platform for Teaching Statistics & Data Analysis While Learning How Students Approach It. https://www.causeweb.org/cause/webinar/teaching/2018-12

[36] Gregory Piatetsky. [n. d.]. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[37] Leo Porter and Daniel Zingaro. 2014. Importance of early performance in CS1: two conflicting assessment stories. In *Proceedings of the 45th ACM technical symposium on Computer science education*. ACM, 295–300.

[38] Anthony Robins. 2010. Learning edge momentum: A new account of outcomes in CS1. *Computer Science Education* 20, 1 (2010), 37–71.

[39] Anthony Robins, Janet Rountree, and Nathan Rountree. 2003. Learning and teaching programming: A review and discussion. *Computer science education* 13, 2 (2003), 137–172.

[40] Edward Sobiesk, Jean Blair, Gregory Conti, Michael Lanham, and Howard Taylor. 2015. Cyber education: a multi-level, multi-discipline approach. In *Proceedings of the 16th Annual Conference on Information Technology Education*. ACM, 43–47.

[41] Svetlana Tishkovskaya and Gillian A Lancaster. 2012. Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education* 20, 2 (2012).

[42] Suleyman Uludag, Murat Karakus, and Stephen W Turner. 2011. Implementing IT0/CS0 with scratch, app inventor forandroid, and lego mindstorms. In *Proceedings of the 2011 conference on Information technology education*. ACM, 183–190.

[43] Fulya Gokalp Yavuz and Mark Daniel Ward. 2018. Fostering Undergraduate Data Science. *The American Statistician* (2018), 1–9.