

# contributed articles

DOI:10.1145/3360646

**A cycle that traces ways to define the landscape of data science.**

BY VICTORIA STODDEN

## The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science

THE EDUCATION AND research enterprise is leveraging opportunities to accelerate science and discovery offered by computational and data-enabled technologies, often broadly referred to as data science. Ten years ago, we wrote that an “accurate image [of a scientific researcher] depicts a computer jockey working at all hours to launch experiments on computer servers.”<sup>8</sup> Since then, the use of data and computation has exploded in academic and industry research, and interest in data science is widespread in universities and institutions. Two key questions emerge for the research enterprise: How to train

the next generation of researchers and scientists in the deeply computational and data-driven research methods and processes they will need and use? and How to support the use of these methods and processes to advance research and discovery across disparate disciplines and, in turn, define data science as a scientific discipline in its own right? An identifiable discipline of data science would encourage and reward research that fosters the continued development of computational and data-enabled methods and their successful integration into research and dissemination pipelines, as well as accelerating the generation of reliable knowledge from data science.

This article offers an intellectual framing to address these two key questions—called the Data Science Life Cycle—intended to aide decision makers in institutions, policy makers and funding agency leadership, as well as data science researchers and curriculum developers. The Data Science Life Cycle introduced here can be used as a framing principle to guide decision making in a variety of educational settings, pointing the way on topics such as: whether to develop new data science courses (and which ones) or rely on existing course offerings or a mix of both; whether to design data science curricula across existing degree granting units or work within them; how to relate new degrees and programmatic initiatives to ongoing research in data science and encourage the development of a recognized research area in data science itself; and

### » **key insights**

- **For Data Science to emerge as a fully fledged science, it is essential to establish intellectual content, ensure knowledge organization, and incorporate external tests of validity for findings.**
- **The Data Science Life Cycle provides a flexible framework that knits stakeholder efforts together to advance Data Science as a science; providing a principled way to include topics such as ethics, reproducibility, and cyberinfrastructure for Data Science, as well as methodological, computational, and domain-specific subjects.**



how to prioritize support for data science research across a variety of disciplinary domains. These can be difficult questions from an implementation point of view since university governance structures typically separate disciplines into effective siloes, with self-contained evaluation, degree-granting, and decision-making authority. Data science presents as a cross cutting methodological effort with the needs of a full-fledged science including: communities for idea sharing, review, and assessment; standards for reproducibility and replicability; journals and/or conferences; vehicles for disciplinary leadership and advancement; an un-

derstanding of its scope; and, broadly agreed-upon core curricula and subjects for training the next generation of researchers and educators.

After motivating the key data science challenges of interdisciplinarity and scope, this article presents the Data Science Life Cycle as a tool to enable the development of data science as a rigorous scientific discipline flexible enough to capitalize on unique institutional strengths and adapt to the needs of different research domains. Examples are given in curriculum development and steps to defining data science as a science.

## Current Approaches to Data Science

There are currently four main approaches taken toward data science at post-secondary institutions and universities in the U.S., with some institutions opting to take more than one approach. The first model involves issuing data science degrees from an existing department or school, such as the computer science department (for example, University of Southern California, Carnegie Mellon University, University of Illinois at Urbana-Champaign), the statistics department (for example, Stanford University), a pro-

fessional studies or extension school (Northwestern University, Harvard University), engineering (Johns Hopkins University), or the School of Information (UC Berkeley). This approach can include innovative steps such as online course offerings or collaborative degrees that approximate data science. An example of the latter is the undergraduate CS+X degree pioneered by the computer science department at the University of Illinois at Urbana-Champaign, where CS refers to computer science and X refers to a domain specific discipline such as economics, anthropology, or linguistics. For a CS+X degree students receive a degree in discipline X with half their courses comprising a common core of computer science classes and half their courses from their disciplinary area X. Stanford University has a CS+X program for undergraduates designed as a joint major between computer science and the humanities. Data science itself has not been established as a sub-discipline in computer science or any other discipline to the best of my knowledge, nor is there an ACM Special Interest Group on Data Science.

The second approach to data science extends or transforms an existing department to explicitly include a home for all of data science, not just the data science degree programs. For example, the statistics department may be renamed Statistics and Data Science (for example, Yale University) or a School of Information Science or Informatics renamed to include the Data Science moniker (Drexel University). The third approach is to create a coordinating mechanism such as a Data Science institute or center at the university (Columbia University, University of Virginia, University of Delaware, University of Chicago, UC Berkeley). Such an institute tends not to have faculty lines, but affiliates faculty who have an appointment elsewhere on campus. It may grant certificates and/or degrees in coordination with affiliated faculty and units, and often began with a focus on professionals and executive education. The University of Washington, for example, extended an existing institute on campus, the eScience Institute, to house its cross-disciplinary Data Science initiative. The final approach is to bring the institute's major

## The Data Science Life Cycle explicitly recognizes the need for data, software, and other artifacts, along with the research findings, to be made available to the community and enables recognition of the need for dedicated research on how this sharing is accomplished.

data science disciplinary units (for example, statistics, computer science and engineering, information science) together under one organizational umbrella to determine degree programs, grant degrees, and house faculty lines and data science research. This is the most recent approach, currently undertaken for example at UC Berkeley (to my knowledge Berkeley is also the only institution to explicitly articulate a Data Science Life Cycle when describing one of its data science degrees).

In some institutions, the trappings of data science have emerged organically within departments themselves without the data science label. For example, offering more classes in statistics and computational methods, or creating data facilities to manage the increasing volumes of data used in departmental research such as the Brain Imaging Data Structure (BIDS) in the Department of Psychology at Stanford University or the Data Analytics and Biostatistics Core in the Emory University School of Medicine. Established domain specific data repositories such as the Protein Data Bank can be central to established research and have long histories of knowledge and expertise development. As data science progresses, we would be remiss not to take the broad advances made by these efforts into account.

It is clear the potential of data science has captured the imagination of students and the broader society.<sup>18</sup> In my experience, however, students can perceive a gap in our pedagogical offerings when it comes to supporting their interest in data science. For a student seeking to do advanced coursework in data science it can appear that statistics is not computational enough, computer science isn't data inference focused enough, information science is too broad, and the domain sciences do not provide a sufficiently deep pedagogical agenda in data science. The research context today is markedly different to even a decade ago in the use of computational and data-enabled methods in a wide range of long-established disciplines from biology (bioinformatics<sup>23</sup>) to physics (computational physics<sup>22</sup>) to mathematics (computer-enabled mathematical proofs<sup>12</sup>) to English

(quantitative analyses of literary texts<sup>13</sup>) to sociology (digital social science<sup>17</sup>), and students are asking the right questions about where data science fits in their education. Not only has it increased the types, scales, and sources of data-accelerated discovery,<sup>25</sup> data has opened new vistas of scientific investigation, methodological advances, and innovation through the creation of novel comprehensive datasets available to communities.<sup>5,16</sup> Data science is inherently interdisciplinary, yet must have a coherent scope in order to develop as a discipline.

### Defining Data Science as a Discipline: The Challenges of Interdisciplinarity and Scope

In what institutional unit or entity should a data science program reside, and what subject matter is considered within the scope of data science? These questions belie the two principal challenges to the advancement of data science as a discipline: its inherently interdisciplinary nature, and the lack of a well-defined scope.

**Challenge 1. Data science is inherently interdisciplinary.** Data science is emergent from a plurality of disciplines, a fact that has been widely noted.<sup>28</sup> These disciplines often exist in different parts of the institution, potentially posing coordination and implementation challenges both within the institution and for data science as an emerging field of research. Few would dispute the central role of data inference methods or software development in data science, yet even those two examples have different loci within the institutional structure: the former typically in a Department of Statistics (often situated in the Faculty of Arts and Sciences) and the latter in computer science departments (often located in the School of Engineering). In addition, schools of information science contribute expertise in data discovery, storage and retrieval, stewardship, archiving, and artifact reuse; engineering and the physical sciences disciplines perform deeply computational simulation-based research; and business schools advance business intelligence and carry out data analytics. The list of examples goes on. These disciplines contribute different but

necessary aspects of a data science discipline and many of the skills used in data science already exist in established departments.

**Challenge 2: Data science must have a well-defined scope.** Many definitions of data science have been put forward, indeed this publication presented its own in 2013: “Data science [involves] data and, by extension, statistics, or the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference” or, “Data science is the study of the generalizable extraction of knowledge from data.”<sup>6</sup> Through conversations in 2013, the following definition was developed by Iain Johnstone, Peter Bickel, Bin Yu, and myself: “Data Science is the science of (collaboratively) generating, acquiring, managing, analyzing, carrying out inference, and reporting on data.” This broad scope means that data science covers a large proportion of the research carried out in institutions today, and implementations of data science programs can be markedly different at different institutions.<sup>20</sup>

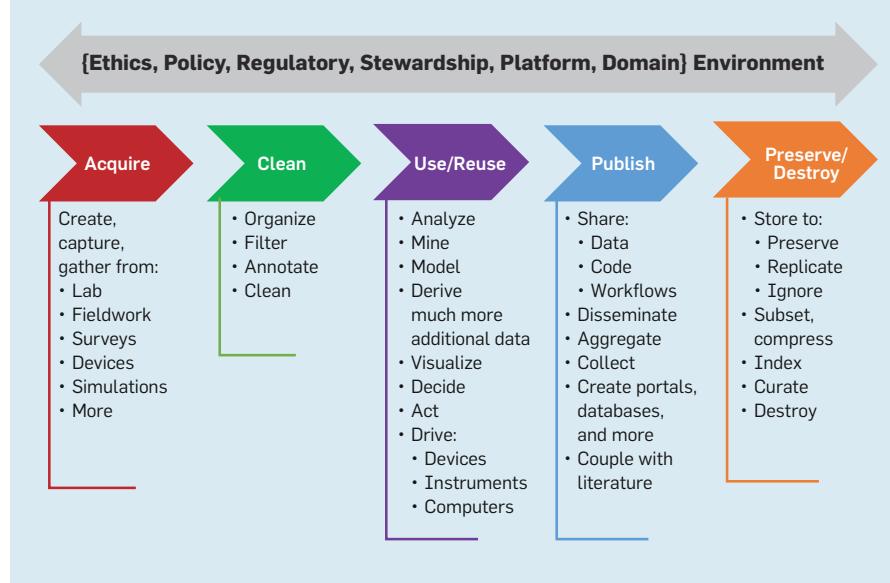
### A Framing for Data Science: The Data Science Life Cycle

Although the Data Science Life Cycle is a new concept, it is an extension of “the Data Life Cycle,” which has a long history in the information sciences and many domain sciences.<sup>1</sup> The Data

Life Cycle describes the various stages a dataset traverses as it undergoes scientific collection and investigation and is typically used to guide data management decisions and practices. I extend this idea beyond its focus on data to describe the complete process of data science with the Data Science Life Cycle. This work extends research in the Data Life Cycle by focusing on the generation of scientific findings, and thereby including computational components, inferential methodology, and articulating a clear role for ethics and meta research within the scope of data science. It can also provide a foundational grounding for data science pedagogical program design.

**Extending the concept of the data life cycle.** Figure 1 shows a depiction of a Data Life Cycle, following a dataset from acquisition, through cleaning, use, publication of the resulting dataset, and then through to an eventual preserve/destroy decision for the dataset. It is important to note that there is no single fixed definition of a Data Life Cycle, rather it’s a thematic abstraction whose manifestation may change depending on the specific dataset or collection of datasets to which it is applied and the purpose of the data collection. A Data Science Life Cycle expands the area of focus beyond the dataset, to the complete bundle of artifacts (for example, data, code, workflow and computational environment information)

**Figure 1. Example of a data life cycle and surrounding data ecosystem (reprinted with permission).**



and knowledge (scientific results) produced in the course of data science research results.

Figure 2 shows a depiction of a Data Science Life Cycle describing stages of data science research, extending the Data Life Cycle reprinted in Figure 1. As in Figure 1, Figure 2 depicts an abstraction, intended to be customized to particular data science projects.

The act of scientific discovery in data science produces findings just like any area of research, and typically creates or leverages other artifacts as well, for example, the data used to support the findings and the code that produces the findings from the data (it may even produce other artifacts as well, for example, curriculum materials, software tools, and hardware prototypes). Research findings and artifacts are viewed with dissemination to

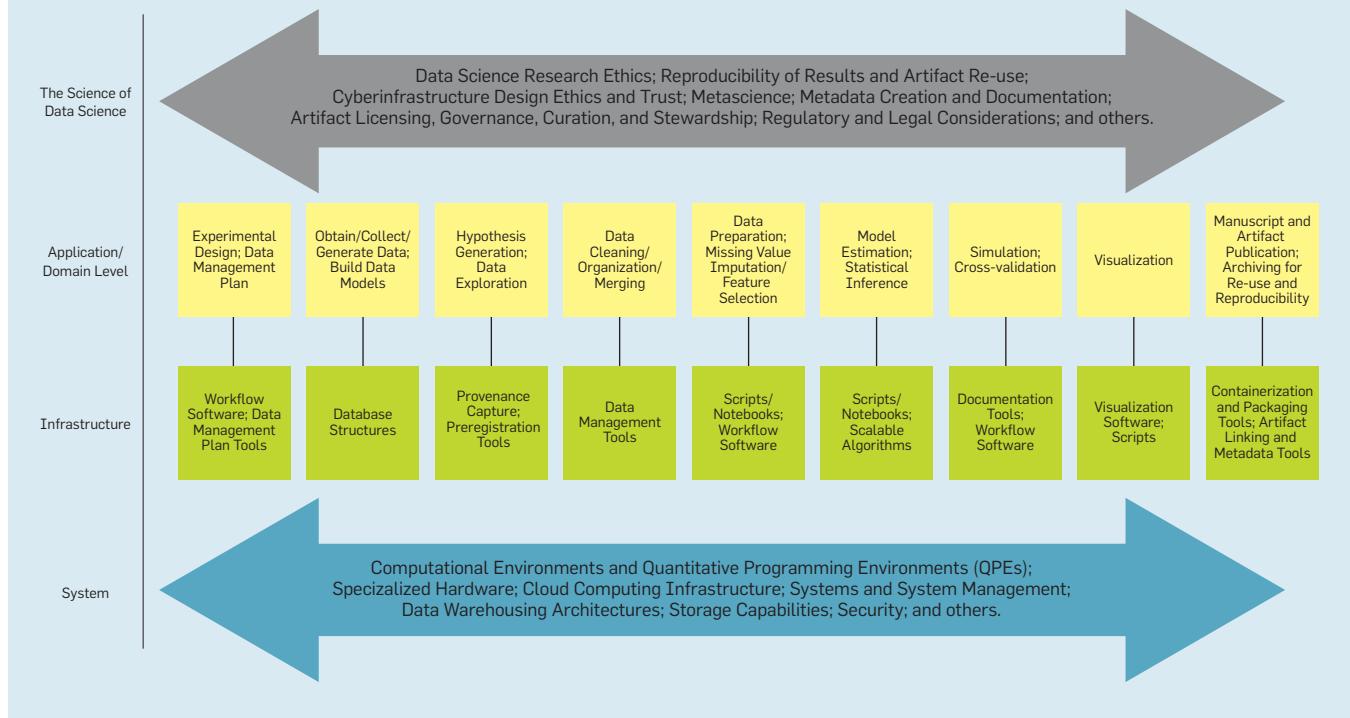
the research community at the point of publication when created. This is what is meant by the term “life cycle”—an explicit recognition that artifacts pass to the community at the point of publication, readied to begin the life cycle again in a new research effort, as inputs. The Data Science Life Cycle explicitly recognizes the need for data, software, and other artifacts, along with the research findings, to be made available to the community and enables recognition of the need for dedicated research on how this sharing is accomplished.

“Reproducibility of Results and Artifact Reuse” is listed as a topic in the overarching grey arrow in Figure 2. The life cycle approach allows a principled incorporation of the notion of computational reproducibility—the practice of ensuring artifacts and

computational information needed to regenerate computational results is openly available post-publication.<sup>4,15,28</sup> Figure 2 emphasizes that artifact preservation activities occur both before and during computation, for the duration of the discovery process. An attempt to recreate computational and data manipulation steps for preservation purposes after publication can be difficult and time consuming, if not impossible. The Data Management Plan, required by the National Science Foundation and other science funders, is therefore included at the beginning of the Data Science Life Cycle, to emphasize the importance of early planning for the artifact preservation that will occur at the point of eventual publication (of the results as well as the supporting artifacts). The need for improved tools for documen-

**Figure 2. An example of a Data Science Life Cycle.**

The light-yellow layer is the Application/Domain Science level. This level portrays the steps of a research project described at the domain level. The green layer beneath is the Infrastructure level, describing the computational infrastructure that enables the Application level. The blue arrow underlying both the application and infrastructure levels is the System level, describing system elements upon which the data science discovery process depends. Similarly, the overarching grey arrow is the meta-scientific level “The Science of Data Science.” The arrows are intended to depict the life cycle: that the output of data science discovery (the findings and the artifacts that enable reproducibility such as code, data, and workflow and computational environment information for example) are published for verification and reuse by others. The figure provides a way to consider tool use explicitly for each discovery step. Some tools may support more than one discovery step, for example, notebooks, and some application level steps may use more than one tool. Different steps in the discovery pipeline, whether at the domain, infrastructure, system, or science of data science level, may be carried out by different people.



tation and recording of the steps in the data science discovery process becomes evident with this approach as does greater recognition that the production of reusable research artifacts (for example, data, software that support a published scientific finding) is a valuable researcher activity.

**Computational and meta-scientific aspects of data science must be explicitly considered.** Crucially, the Data Science Life Cycle adds an additional dimension to the Data Life Cycle: the computational layer that enables data science research. A data scientist may proceed through the steps depicted in the Data Science Life Cycle in Figure 2: experimental design; obtaining/generating/collecting data; data exploration and hypothesis generation; data cleaning, merging, and organization; feature selection and data preparation; model estimation and statistical inference; simulation and cross-validation, visualization; publication and artifact preservation/archiving. This series of steps is called the “Application Level” (depicted in pale yellow in Figure 2), referring to the scientific application or domain of research. As noted, the Data Science Life Cycle is an abstraction and any particular research project may include a subset of these steps.

There are additional components beyond the Application Level in every data science project, depicted by the grey arrow across the top of Figure 2 mentioned earlier, including data science ethics; documentation of the research and meta data creation; reproducibility; and policy and legal aspects including governance, privacy, and intellectual property considerations.<sup>26</sup> This is the “Science of Data Science Level.” In addition, data science projects encompass computational skills and technologies (for example, interpreted languages such as R and Python, data querying languages, distributed computing resources) represented in the green, lower layer, called the “Infrastructure Level” of the Data Science Life Cycle. None of the technologies listed in Figure 2 are prescriptive but they support the steps in the Data Science Life Cycle, in particular the Application Level. Importantly, each are research areas of research and development in their own right, including

## A life cycle approach encourages and enables a unification of views regarding data science and gives us a footing from which to adapt and evolve the practice and teaching of data science to research projects and to institutional strengths.

notebooks and workflow software; visualization tools; statistical inference languages; data management tools; and archiving and artifact linking tools. Running across the entire Data Science Life Cycle, and depicted in the blue arrow at the bottom of Figure 2, are the hardware and other technological structures on which the data science experiment is carried out, including compute infrastructure, cloud computing systems, data structures, storage capabilities, and quantitative programming environments (QPEs).<sup>9</sup> This is called the System Level. Computational reproducibility is an important factor when deciding which artifacts and details in the discovery process to preserve and share. For example, information on how and why parameters were selected in model selection could be included in the documentation and workflow information. The Data Science Life Cycle highlights the various contributions made to the research by different people and could help indicate ways to give appropriate credit by including information on who has contributed what to the discovery process.

**Two simplified examples of the Data Science Life Cycle in research settings.** Here, I present two applications of the Data Science Life Cycle to simplified but representative descriptions of research that illustrate how this approach can surface nuanced and important aspects of data science in different settings. In the first example researchers wish to classify two types of cancer using gene expression data.<sup>10-11</sup> The steps the authors describe for an experiment are as follows:

1. Obtain gene expression data (the data are already split into train/test subsets based on clinical conditions).
2. Normalize the data (including both train/test subsets).
3. Apply Recursive Feature Elimination:
  - a. Train classifier using Support Vector Machines (SVMs).
  - b. Compute a ranking criterion for each feature.
  - c. Remove features with the smallest ranking criteria.
  - d. Iterate until a tolerance threshold is reached.
4. Perform cross-tests with the baseline method from Golub et al.<sup>10</sup> to compare gene sets and classifiers.

Mapping this experimental description to the Application Layer of the Data Science Life Cycle could proceed as follows: Obtain Data → Data Preparation → Feature Selection/Model Estimation → Cross-tests and Validation → Publication and Archiving. Information regarding the tools and software used for each step is then mapped to the Infrastructure Layer and overarching issues, such as data governance and sharing policies, detailed in the Science of Data Science Level. Notice this data science pipeline incorporates a cyclical loop in the pipeline when Recursive Feature Elimination is employed.

The second example gives a stylized description of hypothesis-driven research experiment to test whether a journal's impact factor is related to the existence of a data or code sharing author policy.<sup>27</sup> The steps are as follows:

1. Determine the hypothesis to test.
2. Design an appropriate experiment to test the hypothesis.
3. Collected data on journal impact factors and artifact policies as well as other descriptive information.
4. Test the hypothesis.
5. Report the results.

We map the steps to the Data Science Life Cycle as follows: Determine Hypothesis → Experimental Design → Collect Data → Statistical Inference → Publication. Computational tools used in each step can be detailed in the Infrastructure Level description, and issues that apply to the entire life cycle considered in the Science of Data Science Level, such as data and code availability, preregistration of hypothesis tests, Institutional Review Board (IRB) information, if relevant. Although simplified, these two examples represent different research questions and two different instantiations of the Data Science Life Cycle, but both show how the Data Science Life Cycle framework allows important aspects of the research, such as computational implementations and data ethics, to be cogently and deliberately incorporated as part of the research and publication process.

These examples also illustrate how the Data Science Life Cycle tests whether a particular research effort fits under the rubric of data science. Gaps at the Infrastructure or System Levels can be

## Data science is benefitting from close association with industry as computer science did at its inception.

more easily detected and recognized as part of a comprehensive Data Science research agenda, including for example algorithms; containerization technologies; abstractions of data manipulations; data structures; distributed computing; parallel, cloud or edge computing; hardware design (for example, application specific integrated circuits and their development such as TPUs, or networking capabilities for data distribution).

Considering the Data Science Life Cycle as a life cycle enables a natural consideration of crucial overarching factors such as reproducibility, documentation and meta data, ethics, and archiving of research artifacts such as data and code. The Data Science Life Cycle provides guidance on the multi-faceted set of skills and personnel needed for data science, for example “skills for dealing with organizational artifacts of large-scale cluster computing. The new skills cope with severe new constraints on algorithms posed by the multiprocessor/networked world.”<sup>7</sup> Workforce development is therefore incorporated into the life cycle approach, which is especially germane to data science as “enthusiasm feeds on the notable successes scored in the last decade by brand-name global information technology (IT) enterprises, such as Google and Amazon.”<sup>7</sup>

The Data Science Life Cycle engages relevant stakeholders in the larger research community in a systematic way, including not only data science researchers but others such as archivists, libraries and librarians, legal experts, publishers, funding agencies, and scientific societies. It gives a framework to clarify how different contributions knit together to support each other to advance data science.

### Leveraging the Data Science Life Cycle

A life cycle approach encourages and enables a unification of views regarding data science and gives us a footing from which to adapt and evolve the practice and teaching of data science to research projects and to institutional strengths. There are commonalities to nearly all data science efforts, for example, data wrangling, data inference, code writing, artifact creation and sharing. A common intellectual framework

can facilitate knowledge sharing about data science as a discipline across different fields and domains using data science methods in their research.

**A data science curriculum.** Conceptualizing data science as a life cycle also gives a way to position classes and sequences to teach core and elective data science skills, indicating where existing courses may fit and where new courses may need to be developed. It helps define a curriculum by using the steps of the Data Science Life Cycle as a pedagogical sequence and provides for the inclusion of overarching topics such as data science ethics, and intellectual property, reproducibility, or data governance considerations.<sup>24</sup> Perhaps most importantly the Data Science Life Cycle can indicate courses that may be out of scope and new course topics essential to data science.

The accompanying table shows how several commonly offered courses could be matched to the steps described by the Data Science Life Cycle described in Figure 2. Although not included in the table, each step can be augmented by the creation of new targeted classes if needed, such as Data Policy, Reproducibility in Data Science, Data Science Ethics, Circuit Design for Deep Learning, Software Engineering Principles for Data Science, Mathematics for Data Science, Interoperability and Integration of Different Data Sources, Data Science with Streaming Data, Software Preservation and Archiving, Workflow Tools for Data Science, Intellectual Property for Scientific Code and Data. The list goes on. The addition of domain specific optional courses could define tracks or specializations within a data science curriculum (for example, Earth sciences, bioinformatics, sociology; cyberinfrastructure for data science) to create a potential DS+X degree in the spirit of the CS+X degrees discussed previously.

The emergence of a discipline of data science is necessary to advance data science as well as encourage reliable and reproducible discoveries, elevating the endeavor to a branch of the scientific method. Data science may eventually develop as a set of discipline-adapted discovery techniques and practices, perhaps including a cross-disciplinary core. Data science is benefitting from close association with industry as

computer science did at its inception, for example, IBM's creation of the Watson Scientific Computing Laboratory at Columbia University in 1945.<sup>14</sup> Analysis of consumer data by Google, Facebook, and Amazon is generating prominent successes in image identification and voice transcription among other areas. Opportunities for industry employment and workforce development create an attractive feature of data science at the institutional level.

**Elevating the practice of data science to a science.** The Data Science Life Cycle framework is an essential conceptualization in the development of data science as a science. A recent National Academies of Sciences, Engineering, and Medicine consensus report on “Reproducibility and Replication in Science” spotlights the need to better develop scientific underpinnings for computationally and data-enabled research investigations<sup>21</sup> and a March

#### An example mapping from some routinely offered courses to the steps of the Data Science Life Cycle.

The table is not intended as a complete and comprehensive description of all skills required to be an effective data scientist, but an illustration of how current courses could be incorporated into a data science training curriculum, within which students may pursue pathways of interest. Possible new courses to be developed can be gleaned from such a presentation. Some courses are listed in more than one step to illustrate various ways they might be included in curriculum design.

#### Data Science Life Cycle

Step	Possible (Existing) Courses
Experimental design	<ul style="list-style-type: none"> <li>▶ Introduction to Probability</li> <li>▶ Introduction to Statistics</li> <li>▶ Design of Experiments (including Human Subjects and Informed Consent)</li> </ul>
Obtaining data	<ul style="list-style-type: none"> <li>▶ Experimental Methodology</li> <li>▶ Introduction to Databases</li> <li>▶ Introduction to SQL, noSQL</li> <li>▶ Sensor Integration and Control</li> </ul>
Data exploration	<ul style="list-style-type: none"> <li>▶ Introduction to R</li> <li>▶ Introduction to python</li> <li>▶ Graphics and Data Visualization</li> <li>▶ Introduction to Statistics</li> </ul>
Databases and data structures including cleaning/organizing	<ul style="list-style-type: none"> <li>▶ Introduction to Database Systems</li> <li>▶ Introduction to SQL, noSQL</li> <li>▶ Natural Language Processing (NLP)</li> </ul>
Software engineering	<ul style="list-style-type: none"> <li>▶ Python, R, C, C++, Julia</li> <li>▶ Distributed Systems, MapReduce</li> <li>▶ Software Testing</li> </ul>
Feature selection	<ul style="list-style-type: none"> <li>▶ Statistical Learning</li> <li>▶ Domain-specific courses, for example, Bioinformatics for Transcriptomics; Brain Imaging in Cognitive Neuroscience Research</li> </ul>
Model estimation	<ul style="list-style-type: none"> <li>▶ Mathematics (Probability, Linear Algebra, Calculus, Real Analysis)</li> <li>▶ Applied Statistics</li> <li>▶ Machine Learning</li> <li>▶ Data Mining</li> <li>▶ Deep Learning</li> <li>▶ Scalable Algorithms</li> <li>▶ Statistical Decision Theory</li> </ul>
Simulation and cross-validation	<ul style="list-style-type: none"> <li>▶ Fundamentals of Numerical Methods</li> <li>▶ Introduction to Computer Modeling and Simulation</li> <li>▶ Statistical Learning</li> </ul>
Visualization	<ul style="list-style-type: none"> <li>▶ Information Visualization</li> <li>▶ Scientific Visualization and Graphics</li> <li>▶ [Domain specific courses such as Learning ArcGIS; Spatial Data Visualization]</li> </ul>
Publication/Archiving	<ul style="list-style-type: none"> <li>▶ Introduction to Information</li> <li>▶ Data Archiving and FAIR Data</li> <li>▶ Scientific Report Writing</li> <li>▶ Research Data Management</li> <li>▶ Open Access and Scholarly Communication</li> <li>▶ Digital Libraries and Preservation</li> </ul>
Overarching topics	<ul style="list-style-type: none"> <li>▶ Ethics for Scientists</li> <li>▶ Data Privacy</li> <li>▶ National and International Regulatory Trends in Data Protection</li> </ul>

2019 National Academy of Sciences Colloquium entitled “The Science of Deep Learning” aimed to bring scientific foundations to the fore of the deep learning research agenda.<sup>19</sup> The discussion regarding the scientific underpinnings of data analysis began in 1962, when John Tukey presented three criteria a discipline ought to meet in order to be considered a science:<sup>20</sup>

1. Intellectual content.
2. Organization into an understandable form.
3. Reliance upon the test of experience as the ultimate standard of validity.

If one accepts these criteria, the Data Science Life Cycle can be leveraged to demonstrate intellectual content, promote its organization (see Figure 2), and incorporate external tests of the validity of findings. On this last point, the structure of the Data Science Life Cycle builds in reproducibility, reuse, and verification of results with its embedded notion that artifacts supporting the claims (such as data, code, workflow information) be made available as part of the publication (life cycle) process. Research on platforms and infrastructure for data science facilitates Tukey’s second criterion by advancing organizational topics such as artifact meta data; containerization, packaging and dissemination standards; and community expectations regarding FAIR (findability, accessibility, interoperability, and reusability), archiving, and persistence of the artifacts produced by data science. These efforts also help enable comparisons of data science pipelines to increase understanding of any differences in outcomes of “tests of experience.”<sup>21</sup> The Data Science Life Cycle exposes these topics as areas for research within the discipline of data science.<sup>2</sup> Several conferences and journals have begun to require artifact availability and infrastructure projects are emerging to support reproducibility across the data science discovery pipeline.<sup>3</sup> Considering these issues through a Data Science Life Cycle gives a frame for their inclusion as research areas integral to the discipline of Data Science. Data science without a unifying framework risks being a set of disparate computational activities in various scientific domains, rather than a coherent field of inquiry producing reliable reproducible knowledge.

## Conclusion

Without a flexible yet unified overarching framework we risk missing opportunities for discovering and addressing research issues within data science and training students in effective scientific methodologies for reliable and transparent data-enabled discovery. Data science brings new research topics, for example, computational reproducibility; ethics in data science; cyberinfrastructure and tools for data science. Without the Data Science Life Cycle approach, we risk an implementation of data science that too closely hews to a view that reflects the perspective of a particular discipline and could miss opportunities to share knowledge on data science research and teaching broadly across disciplines. In addition, a Data Science Life Cycle approach can give university leadership a framework to leverage their existing resources on campus as they strategize support for a cross-disciplinary data science curriculum and research agenda. The life cycle approach allows data science research and curriculum efforts to support the development of a scientific discipline, enabling progress toward fulfilling Tukey’s three criteria for a science. ■

## References

1. Berman, F. et al. Realizing the potential of data science. *Commun. ACM* 61, 4, (Apr. 2018), 67–72; <https://cacm.acm.org/magazines/2018/4/226372-realizing-the-potential-of-data-science/fulltext>
2. Bernau, C. et al. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* 30, 12, (2014); <https://academic.oup.com/bioinformatics/article/30/12/I05/388164>
3. Brinckman, A. et al. Computing environments for reproducibility: Capturing the ‘whole tale.’ *Future Generation Computer System* 94, 854–867;
4. <https://www.sciencedirect.com/science/article/pii/S0167739X173010695>
5. Collberg, C. and Proebsting, T.A. Repeatability in computer systems research. *Commun. ACM* 59, 3 (Mar. 2016), 62–69; <https://doi.org/10.1145/2812803>
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 2009; <https://ieeexplore.ieee.org/document/5206848>
7. Dhar, V. Data science and prediction. *Commun. ACM* 56, 12 (Dec. 2013); 64–73; <https://doi.org/10.1145/2500499>
8. Donoho, D.L. 50 years of data science. *J. Computational and Graphical Statistics* 26, 4 (2017); <https://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1384734>
9. Donoho, D.L., Maleki, A., Ur Rahman, I., Shahram, M., and Stodden, V. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* 11, 1, (Jan.-Feb 2009).
10. Donoho, D.L. and Stodden, V. 2015. Reproducible research in the mathematical sciences. J. Higham, ed. *The Princeton Companion to Applied Mathematics*.
11. Golub, T.R. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 5439 (1999), 531–537.
12. Guyon, I. et al. Gene selection for cancer classification using support vector machines. *Machine Learning* 46 (Jan. 2002); <https://doi.org/10.1023/A:1012487302797>
13. Hales, T. Mathematics in the age of the Turing machine. *Turing’s Legacy Developments from Turing’s Ideas in Logic*. R. Downey, ed., 2014; <https://www.cambridge.org/core/books/turings-legacy/mathematics-in-the-age-of-the-turing-machine/376464C81D16F9323EEFB2A2A924D2F4>
14. Hoover, H. Quantitative analysis and literary studies. *A Companion to Digital Literary Studies*. S. Schreibman and R. Siemens, eds. Blackwell, Oxford, U.K., 2008.
15. IBM. The Origins of Computer Science; <https://www.ibm.com/ibm/history/ibm100/us/en/icons/compsci/>
16. Ivie, P. and Thain, D. Reproducibility in scientific computing. *ACM Comput. Surv.* 51, 3 (2018), Art. 63; <https://doi.org/10.1145/3186266>
17. Krizhevsky, A., Sutskever, I., and Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25, 2012. F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds; <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
18. Lazer, D. et al. Computational social science. *Science* 323, 5915 (2009); <http://science.sciencemag.org/content/323/5915/721>
19. Manyika, J. et al. Big Data: The Next Frontier for Innovation, Competition and Productivity. McKinsey Global Institute, 2011; <http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation>
20. NAS Sackler Colloquium. The Science of Deep Learning, 2019; <http://www.cvent.com/events/the-science-of-deep-learning/event-summary-a96a8734faf841ea8d5439e081b50f54.aspx>
21. National Academies of Sciences, Engineering, and Medicine. *Data Science for Undergraduates: Opportunities and Options*. The National Academies Press, Washington, D.C.; <https://doi.org/10.17226/25104>.
22. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, D.C., 2019; <https://doi.org/10.17226/25303>
23. Steering Committee on Computational Physics. *Computation as a Tool for Discovery in Physics*. Report to the National Science Foundation, 2002; <https://www.nsf.gov/pubs/2002/nsf02176/nsf02176.pdf>
24. Ouzounis, C.A. Rise and demise of bioinformatics? Promise and progress. *PLoS Comput Biol* 8, 4 (2012), e1002487; <https://doi.org/10.1371/journal.pcbi.1002487>
25. Saltz, J.S., Dewar, N.I., Heckman, and R. Key concepts for a data science ethics curriculum. In *Proceedings of the 49th ACM Technical Symp. Computer Science Education*. ACM, New York, NY, 952–957; <https://doi.org/10.1145/3159450.3159483>
26. Siewert, S. Big data in the cloud: Data velocity, volume, variety, veracity. *IBM Developer*, July 9, 2013; <https://www.ibm.com/developerworks/library/bd-bigdatacloud/index.html>
27. Stodden, V. The legal framework for reproducible research in the sciences: Licensing and copyright. *Computing in Science and Engineering* 11, 1 (2009), 35–40.
28. Stodden, V., Guo, P. and Ma, Z. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE* 8, 6 (2013), e67111; <https://doi.org/10.1371/journal.pone.0067111>
29. Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., Taufer, M. Enhancing Reproducibility for Computational Methods. *Science* 354, 6317 (Dec. 9, 2016).
30. Stodden, V., Wu, X. and Sochat, V. AIM: An abstraction for improving machine learning prediction. In *Proceedings of the IEEE Data Science Workshop*. (Lausanne, Switzerland, 2018), 1–5.
31. Tukey, J.W. The Future of Data Analysis. *Ann. Math. Statist.* 33, 1 (1962), 1–67.

**Victoria Stodden** (vcs@stodden.net) is a statistician and associate professor at the University of Illinois at Urbana-Champaign, IL, USA.

This material is based upon work supported by National Science Foundation Award #1941443.

Copyright held by author/owner.