# The Effectiveness of Data Science as a means to achieve Proficiency in Scientific Literacy

Wendy Ceccucci
wendy.ceccucci@quinnipiac.edu
Computer Information Systems
Quinnipiac University
Hamden, CT 06518 USA

Dawn Tamarkin
tarmarkin@stcc.edu
Biology
Springfield Technical Community College
Springfield, MA 01105 USA

Kiku Jones
kiku.jones@quinnipiac.edu
Computer Systems Management
Quinnipiac University
Hamden, CT 06518 USA

## Abstract

Data Science courses are becoming more prevalent in recent years. Increasingly more universities are offering individual courses and majors in the field of Data Science. This study evaluates data science education as a means to become proficient in scientific literacy. The results demonstrate how the educational goals of a Data Science course meet the scientific literacy criteria in regards to the process of science. Based on the commonality between data science and scientific literacy courses, the paper concludes that a data science course can be used as an alternative way for students in any major to gain scientific literacy skills.

**Keywords:** Data Science, Scientific Literacy, Scientific Process

### 1. INTRODUCTION

The amount of data produced across the globe has been increasing exponentially and continues to grow. Effectively analyzing these huge collections of data, now called Big Data, can create significant value, increasing competitiveness and delivering more value to consumers. Data science is the general analysis of Big Data. It is the comprehensive understanding of where data comes from, what data represents, and how data can be transformed into meaningful information that can be used to solve problems in diverse domains. It encompasses statistics, hypothesis testing, predictive modeling, understanding the effects of performing computations on data, and how to represent the data to others.

The goal of this paper is to study the effectiveness of data science and visualization as a means to achieve scientific literacy. By utilizing data science techniques, can students

acquire competency in the area of scientific literacy? Some universities are now offering data science courses, and at least one university is now offering data science as a non-lab science course (Squire, 2012). In order to examine how effective data science can be for scientific literacy, an analysis of the learning objectives, educational goals and methodologies used in the introductory data science courses is compared to the objectives of courses that fulfill a scientific literacy requirement.

## 2. BACKGROUND

### Data Science

There are several similar definitions for data science in the literature. Provost and Fawcett define data science as a "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data." Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data."

The term "data scientist" was originally coined by two data analysts working at LinkedIn and Facebook in 2008 (Davenport & Patil, 2012). While there is no consensus on the definition of data science and data scientists, there are some similarities. An article in Fortune magazine described a data scientist as a person who "helps companies make sense of the massive streams of digital information they collect every day, everything from internally generated sales reports to customer tweets." Another source, Data Scientists (2011), defined data scientists as using technology and skills "to increase awareness, clarity and direction for those working with data… Data scientists don't just present data, data scientist present data with an intelligence awareness of the consequences of presenting that data."

Many institutes of higher education are now offering degrees, certifications or courses in the area of data science. The courses are offered by different departments, including Accounting, Mathematics, Computer Science, and Information Systems. In order to understand the goals and objectives of introductory data science courses better, course syllabi from several universities were examined (Attenburg & Provost, 2012; Blumenstock, 2013; Pfister & Blitzstein, 2014; Schutt & Payel, 2013; Squire, 2012).

While the methodology and prerequisites for the introductory courses vary, there are several similarities in all of the classes. They all focus around the six steps of data science as defined by Davenport:

1. Recognize the problem or question.
2. Review previous findings
3. Model the solution and select the variables
4. Collect the data
5. Analyze the data
6. Present and act on the results. (Davenport, 2012)

In the first two steps, students determine the project scope and develop their questions and hypothesis. They research the topic and data. This step may include narrowing down initial ideas about a larger problem to one that is more defined and approachable. The modeling techniques varied between the classes depending on the level of the students.

The way the data collection step is covered varies based upon the course. Generally, sampling techniques are covered and methodologies for data capture are presented. Different tools are used to capture and mine the data, including R, python, Hadoop, web APIs and google searches. In regards to the data storage step, some of the courses discuss tools for large and small data management and storage, another course simply uses Excel for data storage.

For data preparation, munging, scraping, and/or cleaning is completed to get an informative, manageable data set. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate data items from a dataset. At this point, unstructured data is transformed to structured data.

At the data analysis stage, correlations and conclusions about the data are drawn. Depending on the prerequisites and emphasis of the course, the statistical depth of analysis various from simple linear regression and t-tests to basic machine learning.

The last step, data visualization and translation is the communication of information in a clear and effective way through graphical means. A number of different tools are used at this step, including Panda Visualization Tool, Python, Matplotlib, Microsoft Excel, and R.

**Scientific Literacy**

Scientific literacy is generally valued and acknowledged among educators as a desirable student learning outcome. Scientific literacy has received increasing attention over the years, but there is little consensus on its definition. Its meanings are drawn-out and sometimes contradictory (Laugksch, 2000). According to the National Academy of Science (1996, page 22):

"Scientific literacy means that a person can ask, find, or determine answers to questions derived from curiosity about everyday experiences. It means that a person has the ability to describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions. Scientific literacy implies that a person can identify scientific issues underlying national and local decisions and express positions that are scientifically and technologically informed. A literate citizen should be able to evaluate the quality of scientific information on the basis of its source and the methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately".

In order to achieve scientific literacy as described above, one must be able to understand both basic scientific information as well as the process by which science is carried out. These two aspects to scientific literacy were described by Jon D. Miller in 2007 (Ogunkola, 2013) In other words, it is not enough to memorize basic scientific information to be scientifically literate—one needs to also be able to understand how science is carried out. Each aspect of scientific literacy must be considered in more detail.

Science is the "knowledge about or study of the natural world based on facts learned through experiments and observation." (Merriam-Webster, 2014). A similar definition by the Science Council (2014) is "Science is the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence." The most significant difference between these definitions is the inclusion of social sciences in the Science Council definition. Scientific information needed for scientific literacy would include scientific terminology and concepts, potentially from both the natural and social sciences.

After extensive literature review and surveys of science faculty, Gormally, Brickman and Lutz (2012) defined two major categories of scientific literacy skills "1) skills related to organizing and analyzing the use of methods of inquiry that lead to scientific knowledge, and 2) skills related to organizing, analyzing, and interpreting quantitative data and scientific information" (p. 366). Based on responses from faculty on what skills they considered important for scientific literacy, Gormally and colleagues (2012) consolidated the responses into nine set of skills within the two categories. These skills are primarily related to the process of science.

**The Scientific Process**

The scientific process has been described as a set of scientific method steps. The origin of these steps as educational doctrine (beginning in the late 1800s) came from searching for a more interesting and authentic way to carry out science labs, other than simply following standardized lab procedures (Rudolph, 2005). It also arose as an alternative to rote memorization of scientific facts. One of the first proponents of the scientific method was John Dewey, who emphasized the process of knowledge construction over the knowledge itself. More recently, Rissing (2007) showed that by using the process of science students learned better and "had learned to think for themselves."

There is a range in the number of steps in the scientific method today. Commonly, there are 5 (Simon, et al., 2013; Science Made Simple, 2014) to 10 (Crooks, 1961) steps included, and these steps do not have to occur in order (Tignor, 1961). The following are five steps of the scientific method from a current introductory biology textbook (Simon, et al., 2013):

1. Observation
2. Question
3. Hypothesis
4. Prediction
5. Experiment

To begin the scientific method requires observations and questioning. During the observation step, scientists examine the world they are studying and look for anything of interest. As they observe, they write down

descriptions to put their observations into words. Making observations helps them find a topic for study. By restating their observations as questions, they narrow down their observations to find individual questions to ask (often many questions) based on their observations.

The hypothesis step requires that a scientist choose only one of their questions and restate it as a hypothesis. Since the hypothesis is a testable statement, it is usually phrased such that the scientists' expected outcome is incorporated. This step is often the hardest step for a student in a science lab, for two reasons. First, they find it difficult to choose only one variable (from their questions) for their hypothesis. It takes experience to learn to evaluate each possible component of one's questions separately. Secondly, students worry about committing to a possible outcome that could be wrong. The fact that any hypothesis has value, whether it is supported or refuted, is in contrast to student experiences with assessment.

The prediction step feeds into the experiment step. Prediction is when scientists take the hypothesis and make predictions of how it could be demonstrated to begin to visualize how the hypothesis could be tested. If a scientist cannot make predictions from their hypotheses then they cannot begin to formulate a test for it. The experiment that must be done is readily revealed from the predictions. The experiment step includes both carrying out the experiment and recording the results. Some versions of the scientific method separate these components apart, and some even add a step to spell out that the experiment must be done repeatedly.

Note that another step that is often added at the end of these five steps is a sixth "conclusion" step. This conclusion step is where the scientists, based on the information gathered from the previous step, analyze and share what they discovered. The scientists will need to state whether or not the hypothesis was supported. Regardless of the result, it is during this step that the scientist try to provide meaning to the results and share their interpretation of the data with the scientific community.

## 3. RESULTS

In evaluating Data Science as a Scientific Literacy equivalency, it is evident that it does not always fulfill the criterion that scientific literacy include science knowledge and terminology. Data Science does include the scientific literacy criterion of the process of science, by both relatedness to the scientific method and to the skill set of scientific literacy as defined by Gormally and colleagues (2012). The comparison needed to support that statement is provided in this Results section.

The methodology used in Data Science closely matches the scientific method. First, data scientists must try to make sense of the massive amounts of data. To do this, they will begin by formulating the problem. This is where they determine the *questions* they are trying to answer with the data. Let's take for example a large national retailer with store locations throughout the United States as well as online. The data scientists may have observed that there are spikes in sales at certain times of the year. Based on this *observation*, they may question why there are spikes in sales at particular times of the year. They may even be able at this point to provide an educated guess as to what causes these spikes. This is similar to the *hypothesis* step of the scientific process.

Because there is so much data to comb through, smaller sub-sets are often created to provide a more manageable view of the data. This is done in the data collection step of data science. The retailer has customer, vendor, and transaction data flowing in and being stored 24 hours a day, 7 days a week. In order to make sense of all this data, the data scientists develop smaller sub-sets that may be comprised of regional information, type of product sales/purchases, store front or online, or even smaller by specific location. Breaking down this extensive data into subsets is similar to isolating questions by individual variables. By creating the smaller sub-sets, the data scientist is facilitated in creating appropriate hypotheses with singular variables.

After selecting a data sub-set, the data scientists go through a process of preparing the data prior to analysis. As described above, this is where they will "clean" the data. Part of the process of science is understanding whether data is valid; experimental results may not always be pure, and students should learn how to know what data should be included. From these smaller sub-sets, data scientists can begin to analyze and determine what causes shifts in the data, which is similar to a *prediction* step in the scientific method. For example, in regards to the retail industry, data scientists may find that a smaller sub-set of data points to spikes in sales corresponding with various tourist events in the

location they are reviewing. The data scientist may then postulate that for all locations, spikes in data may be explained by the local events held during the year. All of the actions a student would take to this point would also fit into the first category of scientific literacy "Understand methods of inquiry that lead to scientific knowledge" category of scientific literacy skills (Gormally, *et al.*, 2012).

The next step for the data scientists would be to determine what events are held around the various other retail locations. Data scientists would then run the sales data against the local events data to determine whether or not the correlations and conclusions made with the first set of data is proven true. This could be done for all of the smaller sub-sets of data. This is similar to the *experimentation* step in the scientific method.

Finally, based on the information found in the previous steps, the data scientists would make their conclusions and present their findings. They would use techniques to translate their conclusions in a way that is clear to those who need this information to make decisions (data visualization). For the retailer, this may be showing charts that demonstrate the connection of the local events with the spike in sales data. This step is similar to a *conclusion* step in the scientific method. This second half of actions a student would take would also fit into the second scientific literacy category of "Organize, analyze, and interpret quantitative data and scientific information" category of scientific literacy skills (Gormally, *et al.*, 2012).

## 4. CONCLUSIONS

This example clearly demonstrates that the steps of data science parallel the steps of the scientific method. Students in data science courses are thus exposed to a similar scientific processes as those students taking natural science classes. If the purpose of using the scientific method in the classroom is to get students doing rather than memorizing, data science classes would certainly accomplish that active form of learning as well.

Scientific literacy courses are becoming more prominent in higher education (Hobson, 2008). It has been shown that students encouraged to carry out experiments using the process of science rather than follow step-by-step instructions performed much better (Rissing, 2007); Rissing specified that the improved scores reflected that students using the science process had learned to think for themselves. By using this same process in a data science course students could gain the confidence to think for themselves in other courses as well.

Data science generally deals with data taken from pre-existing data warehouses or marts whereas science courses typically derive their data through experimentation. Both types of science go through the same steps of problem solving. The process of questioning, generating a hypothesis, evaluating, processing, analyzing, and presenting the data are done similarly.

The strong parallels between data science and scientific literacy suggest that a student taking a data science course would gain the same skills of objectivity and analysis as a student in a natural science course. Therefore, whether a data science course is offered in business, computer science, or any other field, that course could fulfill student requirements for scientific literacy. This opens up alternative options for students in any major to gain the important skills of scientific literacy.

## 5. REFERENCES

Attenburg, J. & Provost, F. (2012) NYU INFO-3359 Practical Data Science Syllabus retrieved from: people.stern.nyu.edu/ja1517/pdsfall2012/PDS-syllabus-F12orig.pdf

Blumenstock, J. (2013) University of Washington INFX 598 Introduction to Data Science Syllabus, retrieved from: jblumenstock.com/teaching/course=infx598

Crooks, Kenneth B. M. (1961) Suggestions for Teaching the Scientific Method. *American Biology Teacher*, 23(3), 154-159.

Data Scientist (2011), "What is Data Science" retrieved from: www.datascientists.net/what-is-data-science.

Davenport, T., & Patil, D.J. (2012) Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, October.

Davenport, T. (2013) Keeping up with the Quants. Harvard Business Review Magazine, July-August, retrieved from: http://hbr.org/2013/07/keep-up-with-your-quants/ar/1)

Gormally, C., Brickman, P., & Lutz, M. (2012) Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates' Evaluation of Scientific Information and Arguments. *CBE-Life Sciences Education*, 11(Winter), 364-377.

Hobson,A.(2008). The Surprising Effectiveness of College Scientific Literacy Courses. *The Physics Teacher*, 46, 404-6

Laugksch, R. C., & Spargo, P. E. (1996). Development of a pool of scientific literacy test-items based on selected AAAS literacy goals. *Science Education*, 80(2), 121-143.

Laugksch, Rudiger C. (2000). Scientific literacy: A conceptual overview. *Science Education*, 84(1), 71–94.

Lev-Ram, Michael (2011). Data scientist: The hot new gig in tech. *Fortune*, Sept 5. Retrieved from: tech.fortune.cnn.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech/

Merriam-Webster Online, retrieved from: merriam-webster.com/dictionary/science

Miller, Jon D (2007). The impact of college science courses for non-science majors on adult science literacy," a paper presented to a symposium titled "The Critical Role of College Science Courses for Non-Majors at the annual meeting of the AAAS, 18 Feb. 2007, San Francisco.

National Academy of Science, (1996). National Science Education Standards. National Academy Press, Washington, D.C.

Ogunkola, B. (2013) Scientific Literacy Conceptual Overview, Importance and Strategies for Improvement. *Journal of Educational and Social Research*, 3(1) 265-274.

Pfister, H. & Blitzstein,J. (2014) Harvard School of Engineering and Applied Science CS109 Syllabus. retrieved from: cs109.org/syllabus.php

Provost, F. & Fawcett T., "Data Science and Its Relationship to Big Data and Data-Driven Decision Making", *Big Data*, March 2013, p52. http://online.liebertpub.com/doi/pdf/10.1089/big.2013.1508

Rissing, Steven (2007) Scientific Literacy Happens—when students think for themselves. Taken from http://www.eurekalert.org/pub_releases/2007-02/osu-slh021307.php

Rudolph, John L. (2005) Epistemology for the Masses: The Origins of "The Scientific Method" in American Schools. *History of Education Quarterly*, 45(3), 341-376.

Schutt, R., & Patel, K. (2013) Introduction to Data Science, Columbia University Syllabus, taken from: columbiadatascience.com/about-the-class/about-the-course-2013/

Science staff, (2011), Special Edition Introduction Challenges and Opportunities. Science, 331(6018), 692-693, retrieved from: www.sciencemag.org/site/special/data/

Science Council (2014) Working Collectively to Advance UK Science, Retrieved from: www.sciencecouncil.org/definition

Science Made Simple, (2014) Retrieved from: www.sciencemadesimple.com/scientific_method.html

Simon, Eric J., Dickey, Jean L., Reece, Jane B. (2013) Campbell Essential Biology, 5th edition. Pearson Education, Inc., San Francisco.

Speyer, P. (2014) Six Steps for Applying Data Science: It's all about Teamwork, Health Data Innovation. Retrieved from: www.healthdatainnovation.com/content/six-steps-applying-data-science-its-all-about-teamwork

Squire, Megan (2012) Elon University CSC/ISC 111 Data Science and Visualization Syllabus.

Tignor, Donald M. (1961) The Scientific Method: Another Look. *American Biology Teacher*, 23(3), 160-164.

**Editor's Note:**

*This paper was selected for inclusion in the journal as an ISECON 2014 Meritorious Paper. The acceptance rate is typically 15% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2014.*