DESIGNING A DATA FUNDAMENTALS COURSE

Lynn R. Heinrichs
Department of Computing Sciences
Elon University
Campus Box 2320
Elon, NC 27244
336.278.6192
lheinrichs@elon.edu

ABSTRACT

In today's world, everything can be monitored and measured. Access to the Internet provides an endless source of data that organizations can exploit for business intelligence. The cost of storage has dwindled to pennies. Yet, having data alone is not enough; it must be mined, analyzed, and visualized in order to make sense. In response to the changing data landscape, the author's institution recently made a curricular paradigm shift by transforming its BS in Computer Information Systems to a BS in Information Science. The major in information science provides graduates with data-intensive skills that can be applied across disciplines such as business, biology, criminal justice, and communications. This paper describes the development and implement of a core course in the new program entitled, "Fundamentals of Data."

INTRODUCTION

Faculty members in the information systems field are more than accustomed to the constant challenge of delivering state-of-the-art curricula. Some changes are mostly course maintenance such as updates related to new software releases. Other changes reflect significant paradigm shifts that cause an entire upheaval of a curriculum to reflect a new direction.

The author recently participated in a substantial effort to rechart a BS in Computer Information Systems to a BS in Information Science. The decision to undertake a dramatic change was based upon both external and internal factors. The new data-driven curriculum will be offered beginning Fall 2011 with emphasis on competencies in software, data, and statistics. At the core of the new curriculum is a course entitled "Fundamentals of Data," a prerequisite to several upper-level courses including database management and information security. While there are many models for developing database courses, there are few examples of what concepts and skills should be included in a data fundamentals course. The purpose of this paper is to describe the development and implementation for the first offering of the course in Fall 2011.

THE EXPLOSION OF DATA

According to Redman (2008), information technologies are rapidly becoming commodities that are cheaply available to all. These technologies enable organizations to: (1) acquire and store vast amounts of data and information, (2) deliver data from one place to another, (3) use data for basic organizational transactions, and (4) manipulate data to create new data and information.

As the world we live in becomes more and more interconnected, the amount of data that can be collected and stored has exploded. "Whether we're talking about web server logs, tweet streams, online transaction records, "citizen science," data from sensors, government data, or some other source, the problem isn't finding data, it's figuring out what to do with it" (Loukides, 2010). Data is the raw material of knowledge. In today's world, everything can be monitored and measured. According to Brynjolfsson, an economist and director of the Massachusetts Institute of Technology's Center for Digital Business, "the big problem is going to be the ability of humans to use, analyze and make sense of the data" (Lohr, 2009).

Facilitating the collection and retention of data has been the incredibly cheap cost of storage. In the book, *delete: The Virtue of Forgetting in the Digital Age,* Mayer-Schonberger (2009) describes how storage costs have plummeted over time:

> In 1957, IBM introduced the 305, a computer with magnetic disks as storage devices that offered up to 5 megabyte of space, and which was valued at around $1 million (in 2006 terms). The cost of the storage unit alone ran to about $70,000 per megabyte in the 1950s; by 1980 that price had come down to below $500 (all in 2006 U.S. dollars), less than one percent of what it had been just two-and-a-half decades earlier. Twenty years later, in 2000, storage cost had plummeted to about 1¢, 1/50,000th of what it was in 1980. And in 2008, the cost of storage for one megabyte of information had been reduced to one hundredth of a cent. For fifty years, the cost of storage had roughly been cut in half every two years, while storage density increased 50-million fold . . . (pp. 62-63).

The Information Age is unfolding in three distinct phases (Redman, 2008). The first phase involved the development of a coherent IT infrastructure by organizations; that phase is coming to an end. The second phase involves improving data accuracy by as much as two orders of magnitude. The last phase, exploiting data and information, will be full of tough issues to address.

Organizations are challenged to find ways in which their vast repositories of data can provide competitive advantage. IBM is helping organization's see the benefits and opportunities of data through its Smarter Planet Web site. The site answers the question "why data matters" by sharing insights gained from collaborations with more than 600 organizations worldwide:

> Data is being captured today as never before. It's revealing everything from large and systemic patterns—of global markets, workflows, national infrastructures and natural systems—to the location, temperature, security and condition of every item in a global supply chain. And then there's the growing torrent of information from billions of individuals using social media. They are customers, citizens, students and patients. They are telling us what they think, what they like and want, and what they're witnessing. As important, all this data is far more real-time than ever before (IBM, n.d.).

Mike Loukides (2010), a senior editor for O'Reilly Media, predicts that the "future belongs to the companies who figure out how to collect and use data successfully." This means not just companies using their own data, but also data mashed up from many other sources.

The challenge of turning data into business intelligence or new products creates the need for employees with an appropriate skill set. In an interview with *McKinsey Quarterly*, Hal Varian explained:

> The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids . . . now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it (McKinsey, 2009).

Traditional information systems programs include coursework in database management and other data-based skills. However, the traditional IS program does not really provide the complete skill set required for today's data-driven environments.

DATA-DRIVEN SKILLS AND EMPLOYMENT

What skills are needed to succeed in a data-driven world? Mason and Wiggins (2010) have developed a taxonomy of data science which addresses this question. "Both within the academy and within tech startups, we've been hearing some similar questions lately: Where can I find a good data scientist? What do I need to learn to become a data scientist?" Mason and Wiggins describe the five primary activities of the data scientist as obtain, scrub, explore, model, and iNterpret "(or, if you like, OSEMN, which rhymes with possum)."

Flowingdata.com (2009) describes the emergence of data science as a new field and the skills of the data scientist as blending expertise areas that seem sometimes disjointed:

- Computer Science - acquire and parse data.
- Mathematics, statistics, and data mining - filter and mine.
- Graphic Design - represent and refine.
- Information visualization and human-computer interaction (HCI) – interaction.

The job outlook for information technology positions has remained strong, even in a weak economy. Of the 2010 graduates at the authors' institution in either CIS or computer science, 100% found related employment or enrolled in graduate school. According to the Occupational Employment Outlook for 2008-2018 (BLS, 2010), the job opportunities in computer-related industries will continue to remain strong:

- The Information Sector includes fast-growing, computer-related industries. The data-processing, hosting, and related services industry, are expected to grow by 53 percent; this includes establishments that provide Web and application hosting and streaming services.
- The Professional, Scientific, and Technical Services sector includes employment in computer systems design and related services. It is expected to increase by 45 percent, accounting for nearly one-fourth of all new jobs in this industry sector.

> Employment growth will be driven by increasing demand for the design and integration of sophisticated networks as well as Internet and intranet sites.

Getting a handle on employment growth for data-related jobs is more complex. O'Reilly Research examined job listing data for two open source projects, Hadoop and Cassandra, which are good proxies for the market as a whole in "data-based" employment (Loukides, 2010). A steady year-over-year increase in job listings was evident related to the two open source projects. In addition to O'Reilly's research, the McKinsey 2011 *Big Data* report (McKinsey, 2011) showed that demand for analytical talent in the US could be 60% greater than its projected supply over the next 7 years. The report goes on to explain "The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data. (p. 7)"

Other anecdotal data such as IBM's "Why Data Matters" campaign illustrate the need for professionals with strong data skill sets. Collecting, storing, securing, analyzing, and visualizing the data will be a high priority of leaders in many industries, providing the impetus for job growth in the information and professional services sectors.

A NEW CURRICULUM FOR A DATA-DRIVEN WORLD

To meet the growing need for data professionals, the author's institution recently participated in a substantial effort to rechart a BS in Computer Information Systems to a BS in Information Science. The decision to undertake a dramatic change was based upon both external and internal factors. *Externally*, increased Internet connectivity is giving organizations greater access to endless sources of data that can be mined for business intelligence. A growing need has emerged for professionals who are "data" specialists. *Internally*, the a new curriculum became a natural extension of strategic plan goals to develop partnerships with other programs on campus and to build greater synergy among programs within the department.

The new curriculum is summarized in Table 1. The program centers on using technology to solve a wide-range of complex problems that involve capturing, analyzing, visualizing, and managing large sets of data. Students learn to apply state-of-the-art tools and techniques for transforming a barrage of data to consistent, trusted, and relevant information that can provide insight and support decision making. The hands-on program provides a solid foundation in programming, data, interface design, and statistics. Because of its interdisciplinary nature, information science complements a variety of majors and minors.

Early on in the program, students complete a course in the "Fundamentals of Data" that serves as a prerequisite to both database management and information security. During the curriculum development process, faculty members were not completely sure as to what the content of the course should be. They simply believed that students needed an introductory data course in the same way an IS program might include an introductory programming/application development course. The following course description was developed during the curriculum design process:

ISC 245 Fundamentals of Data. An introduction to the storage, organization, and management of data resources. Topics include data representation, data formats, data files, data storage, and data integrity. Prerequisites: ISC111 or CSC130 or instructor permission.

Table 1. BS in Information Science

Required Courses	**Course Title**	
	CSC/ISC 111 Data Science and Visualization	
	ISC 245 Fundamentals of Data	4 sh
	ISC 301 Database Management and Analysis	4 sh
	ISC 310 Human Computer Interaction	4 sh
	ISC 320 Data Mining and Analytics	4 sh
	ISC 325 Data Driven Web Development	4 sh
	ISC 345 Information Security	4 sh
	ISC 430 Senior ISC Experience	4 sh
Electives	4 hours of upper-level ISC or CSC230	4 sh
Math	MTH 206 Discrete Structures	4 sh
	STS 212 Statistics in Application	4 sh
Computing	STS 327 Statistical Computing	4 sh
	CSC 130 Computer Science I	4 sh
TOTALS	**Total**	**52 sh**

Often when developing a new course, textbooks in the subject matter can be helpful in determining appropriate content and objectives. Unfortunately, the author was unable to identify any current textbooks for this purpose. Internet searches using terms related to "fundamentals" and "data" tend to return titles involving database or statistics concepts. One text that was helpful in providing a starting point was Englander's, *The Architecture of Computer Hardware, Systems Software, and Networking: An Information Technology Approach* (Englander, 2010). Several chapters on the architectural aspects of data including number systems, number representation, and data formats are included as well as file organization and management concepts.

Using these ideas as a starting point, the author developed the following learning objectives for the course:

- Convert and interpret data using different numbering systems.
- Interpret data values for different data formats and standards for text, integer, and real numbers.
- Compare different methods for representing negative numbers.

- Describe methods for storing data using popular multimedia file formats.
- Effectively use a command language for manipulating files and navigating a directory structure.
- Create applications for validating and processing both sequential and random file data.
- Map and convert data between different applications.
- Compare and contrast the use of files versus databases for storing data.

The course begins with the most elemental units of data (bits and bytes), progresses to file organization and management concepts, and positions students to continue their study with databases in a subsequent semester. Students are also prepared with a technical understanding of data necessary for studying information security. The course assumes the introduction of a programming language for exploring file processing concepts, but leaves the specific language choice to the instructor. During the first offering, the author has chosen to use PERL in a Linux environment. The author's interest in PERL stems from previous experience co-teaching a bioinformatics course for non-majors. PERL does not require a lot of overhead in terms of learning time or computer resources. Students can get started very quickly writing their first programs on either a Windows or Mac platform.

FUTURE PLANS

The data fundamentals course is offered only in the fall semester. At the time of this manuscript submission, the first offering of the course is underway. In the spring semester, the two upper-level courses fed by data fundamentals (database and information security) will be offered, and the faculty will have a better sense of how well the courses are working together. The department engages in a formal assessment meeting at the end of the academic year at which time course adaptations can be considered.

REFERENCES

Bureau of Labor Statistics (December 2010). *Occupational outlook handbook 2010-11 edition,* U.S. Department of Labor. Available: http://www.bls.gov/oco/oco2003.htm.

Englander, I. (2010). *The architecture of computer hardware, systems software, and networking: An information technology approach, 4th edition.*

Flowingdata.com (June 4, 2009). *The rise of the data scientist*. Available: http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/.

IBM (n.d.). Welcome to the decade of smart, *Building a Smarter Planet: 1 in a Series.* Available*:* http://www.ibm.com/smarterplanet/global/files/us__en_us__overview__decade_of_smart_011310.pdf.

Lohr, S. (August 5, 2009). For today's graduate, just one word: Statistics, *The New York Times* [Online]. Available: http://www.nytimes.com/2009/08/06/technology/06stats.html.

Loukides, M. (June 2010). *What is data science?* [Online O'Reilly Radar Report]. Available: http://radar.oreilly.com/2010/06/what-is-data-science.html.

Mason, H. and Wiggins, C. (September 25, 2010). *A taxonomy of data science*. Available: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/.

Mayer-Schonberger, V. (2009). *delete: The virtue of forgetting in the digital age*, Princeton University Press, Princeton, NJ.

McKinsey Quarterly (January 2009). Hal Varian on how the Web challenges managers, *McKinsey Quarterly* [Online, Visitor Edition]. Available: https://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286 .

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competitiveness, and productivity*. Available: http://www.mckinsey .com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf

Redman, T. C. (2008). *Data-driven: Profiting from your most important business asset*, Harvard Business School Publishing, Boston, MA.