Situating Computational Thinking with Big Data: Pedagogy and Technology

Austin Cory Bart
Virginia Tech
2202 Kraft Drive
Blacksburg, VA
acbart@vt.edu

ABSTRACT

As Computational Thinking becomes pervasive in the undergraduate curriculum, students must be educated in contexts that are meaningful, authentic, and useful to their long-term needs. I propose working with big data as a novel context for introductory programming, authentic given its strategic importance in diverse fields such as agriculture, history, science, and more. Big data has historically been considered a difficult topic to teach because of the many technical obstacles inhibiting the process. To solve these difficulties, I have introduced a new project: CORGIS–a "Collection of Real-time, Giant, Interesting, Situated Datasets," which encompasses two distinct goals: (1) create a new context for introductory programming, and (2) introduce new content related to big data. The CORGIS project comprises a collection of libraries that provide an interface to big data for students, architectures for rapidly enabling new high velocity and high volume data sources, and a web-based textbook for disseminating relevant course materials. In this paper, I describe the background of this work, the learning theory that it is built upon, my approach, and the ongoing evaluation required to determine its success.

Categories and Subject Descriptors

K.3.2 [**Computer and Information Science Education**]: Computer Science Education

General Terms

Design, Human Factors, Reliability

Keywords

big data, learning enhancement, projects, introductory, volume, velocity

1. MOTIVATION AND RESEARCH PROBLEM

A report by MGI and McKinsey's Business Technology Offices declares that "... by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of Big Data to make effective decisions" [7] . This gap, and the expanding recognition of "Computational Thinking" as a 21st century skill, increasingly requires that computation be positioned in a university's general education curriculum [12]. For instance, Virginia Tech is now formalizing this requirement through a new course ("Introduction to Computational Thinking") that all university students will eventually be required to take [9]. Such a course poses serious pedagogical and motivational challenges: how do we introduce Computational Thinking to students with no prior computing experience and convince them that the field can tangibly benefit their respective disciplines?

An introduction to programming with big data can be uniquely effective at answering this question, as it offers both authentic *context* and course *content.* Few other approaches offer both of these benefits at the same time. Indeed, with respect to content, computational techniques for handling big data have become crucial to interdisciplinary work in science, humanities, agriculture, medicine, and other university subjects [1]. With respect to context, computing educators can make a strong case that leveraging the promise of big data presents unprecedented opportunities for improved economic growth and enhanced innovation. Morever, big data sources can also be found that have a personal connection with students lives – e.g., social media data from students' Facebook page.

Unfortunately, big data is difficult to work with, by its very nature as "quantities of information that cannot be handled with traditional methods" [7], and is usually reserved for advanced courses that non-majors are never intended to reach. In my research, I explore the challenging problem of developing scaffolding and pedagogical technology to enable introductory students to start working with big data immediately. My primary contribution is the CORGIS Datasets project, a "Collection of Real-time, Giant, Interesting, Situated Datasets". The key technical element of this project is a reusable architecture and toolchain for rapidly creating big data libraries. I am also responsible for the instantiation of many libraries to satisfy the diverse array of majors present in the course's offering. Finally, I am developing a new web-based, interactive textbook that leverages contextualization through big data. All of these materials are being used and evaluated in the aforementioned new course for non-majors at Virginia Tech.

2. BACKGROUND AND RELATED WORK

A precise definition for big data is the "3V Model" [5], which posits that there are three dimensions that distinguish big data from ordinary data. First is the volume, the total length (number of records) or breadth (fields within a record) of the information, often measured in bytes. Next is velocity, the rate at which new information is added to the system. Finally is variety, the format(s) of the data, which can vary from well-structured textual formats to complex, diverse encodings (e.g., music data, image data, geospatial data).

These different forms of data bring unique challenges to a classroom. High velocity data is inherently web-based, forcing students to deal unstable internet connections, volatile API interfaces, and data that can come in a variety of forms. It is difficult to deliver high volume data efficiently to students. Both datasets can be hard to find in a format that is readily consumed by novice students using introductory level techniques – requiring conversations about file handling, url access, and many other intermediate to advanced methods.

2.1 A Learning Theory for Big Data

To ground my work in theory, I have drawn upon Situated Learning Theory [4], which argues that learning normally occurs as a function of the activity, context, and culture in which it is situated [6]. Therefore, tasks in the learning environment should parallel real-world tasks, in order to maximize the authenticity of the experience. The use of data analysis as a form of contextualization represents a new and actively growing movement. Recently, low velocity and low volume data sets have been provided by Anderson et al to explore interesting questions in CS-0 courses [2]. The intention of these datasets is to give a uniform experience to the class, ignoring students individual interests and backgrounds – for example, all students will complete an assignment using biological data, and then an assignment using geographic data, and so on. Upper division courses have employed these situated learning experiences using data of varying size and complexity for several years [10, 11]. However, these come late in the curriculum, and offer no scaffolding for beginners. It is unprecedented to use big data in low-level courses.

3. APPROACH AND UNIQUENESS

My previous contribution to this research area was the RealTimeWeb project, a toolchain that provides introductory programming students with an easy way to access high velocity data, also known as "real-time data" [3]. The core of this toolchain is a collection of client libraries through which students can access web-based data. These libraries are available in several different common beginner languages (currently Racket, Python, and Java), and readily available in our online, curated gallery [1]. New libraries can be rapidly generated from a formal specification of the web services' endpoints and data structures through our special, web-based compiler.

I am now expanding RealTimeWeb into the CORGIS Datasets project: the Collection of Real-time, Giant, Interesting, Situated Datasets. The goal of this project is to provide tools and materials for instructors to introduce big data topics to students as early in the curriculum as possible. The focus has expanded from high velocity data to now include support for high volume data. To facilitate this expansion, there is a new subsystem named Eve, which rapidly web-enables well-structured, high volume data sets. By distributing the data set on a high-capacity server, students are no longer responsible for managing the massive files required. Connections to the data can be rapidly produced using the RealTimeWeb's generator system.

In addition to the new technical contributions, I am also developing a web-based, interactive textbook to deliver pedagogical material for students. This system is based on the Runestone platform [8] that provides rich web features such as videos, code execution, and code visualization. I am working closely with the core Runestone developers to integrate CORGIS materials into the interactive coding materials available in the book, along with other features inspired by Situated Learning Theory.

4. RESULTS AND UNIQUENESS

The RealTimeWeb toolchain has been deployed for several semesters in introductory Computer Science courses for majors, ranging from the first course all the way to a Data Structures level course. These integrations ranged from small assignments to entire semester projects using the software. So far, the focus of the evaluation has been on the motivational influence of the system. Quantitative data was collected by surveying students attitudes using well-established motivational frameworks and instruments, and indicates that students tended to find real-time data engaging [3]. In some courses, qualitative data was gathered through small group interviews, where students attribute increased engagement with the authentic, real-world connection offered by real-time data.

As the project is being deployed in the non-majors course on Computational Thinking, I will be expanding the evaluation to additionally cover cognitive gains attributable to the contextualization approach. Surveys and focus groups will be used periodically throughout the semester to more extensively assess not just students engagement, but also their level of understanding of the course concepts, and how that understanding is connected to the use of our materials. Finally, usability studies will be conducted on the interface of the libraries and textbook to determine how successful they are at supporting students programming.

5. CONCLUSION

In this paper, I have described my on-going research on integrating big data into introductory computing courses through scaffolding technology. As a form of both content and contextualization, our results indicate that big data is a unique way to motivate and educate students on an authentic, useful method. I detail the new open-source technologies and materials I have created that make it easy for instructors to integrate big data into their course early and often. Preliminary results from our research indicate that this is a successful approach, and I am planning on conducting more in-depth analysis this fall on a course for non-majors.

6. REFERENCES

[1] `think.cs.vt.edu/corgis`

[1] C. Anderson. The end of theory. *Wired magazine*, 16, 2008.

[2] R. E. Anderson, M. D. Ernst, R. Ordóñez, P. Pham, and S. A. Wolfman. Introductory programming meets the real world: Using real problems and data in cs1. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, pages 465–466, New York, NY, USA, 2014. ACM.

[3] A. C. Bart, E. Tilevich, S. Hall, T. Allevato, and C. A. Shaffer. Transforming introductory computer science projects via real-time web data. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, pages 289–294, New York, NY, USA, 2014. ACM.

[4] M. Ben-Ari. Situated learning in computer science education. *Computer Science Education*, 14(2):85–100, 2004.

[5] L. Douglas. The importance of âĂŸbig dataâĂŹ: A definition. *Gartner (June 2012)*, 2012.

[6] J. Lave and E. Wenger. *Situated learning: Legitimate peripheral participation.* Cambridge university press, 1991.

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[8] B. Miller and D. Ranum. Runestone interactive: Tools for creating interactive course materials. In *Proceedings of the First ACM Conference on Learning Scale Conference*, LS '14, pages 213–214, New York, NY, USA, 2014. ACM.

[9] O. of the Senior Vice President and Provost. Academic implementation strategy for a plan for a new horizon: Envisioning virginia tech 2013-2018. Technical report, 2013.

[10] L. Torgo. *Data Mining with R, learning with case studies.* Chapman and Hall/CRC, 2010.

[11] M. Waldman. Keeping it real: utilizing NYC open data in an introduction to database systems course. *J. Comput. Sci. Coll.*, 28(6):156–161, June 2013.

[12] J. M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006.