Preparing, Visualizing, and Using Real-world Data in Introductory Courses

Austin Cory Bart
Virginia Tech
Blacksburg, Virginia
acbart@vt.edu

Kalpathi Subramanian
The University of North Carolina at Charlotte
Charlotte, North Carolina
krs@uncc.edu

Ruth E. Anderson
University of Washington
Seattle, Washington
rea@cs.washington.edu

Nadeem Abdul Hamid
Berry College
Mount Berry, Georgia
nadeem@acm.org

CCS CONCEPTS

• **Information systems** → **Information integration**; Data management systems; • **Social and professional topics** → *Model curricula*; Computing literacy;

KEYWORDS

Real world data, data science, visualization

ACM Reference format:
Austin Cory Bart, Kalpathi Subramanian, Ruth E. Anderson, and Nadeem Abdul Hamid. 2018. Preparing, Visualizing, and Using Real-world Data in Introductory Courses. In *Proceedings of SIGCSE '18: The 49th ACM Technical Symposium on Computer Science Education, Baltimore , MD, USA, February 21–24, 2018 (SIGCSE '18),* 2 pages.
https://doi.org/10.1145/3159450.3159616

1 SUMMARY

Working with real-world data has increasingly become a popular context for introductory computing courses [1, 5, 7, 9, 10]. As a valuable 21st century skill, preparing students to be able to divine meaning from data can be useful to their long-term careers [4]. Because Data Science aligns so closely with computing, many of the topics and problems it affords as a context can support the core learning objectives in introductory computing classes. In many instances, incorporating a real-world dataset to provide concrete context for an activity or assignment can improve student engagement and understanding of the abstract educational content being presented.

However, there are many problems inherent to bringing real-world data into introductory courses. How do instructors, with finite amounts of time and energy, find and prepare suitable datasets for their pedagogical needs? Once the datasets are ready, how can students conveniently interact with and draw meaning from the datasets, especially when they are used in complex projects that are typical of later introductory courses? On the other hand, how does an instructor balance the complexities of using real-world datasets in the classroom, making sure that students appreciate the meaningfulness of course activities and their connection to learning objectives?

This panel brings together experts with experience in using real-world data in introductory computing courses. Each panelist provides unique perspectives and skills to the problem of preparing, interacting, visualizing, and using pedagogical datasets. This panel should be of particular interest to instructors who are considering integrating current and real-world data into their assignments and projects, and to educational developers who want to create and manage datasets for pedagogical purposes. The panel will follow a conventional format: 5 minutes of introduction, 10 minutes for each panelist to present, and then 30 minutes for audience Q&A.

2 AUSTIN CORY BART(MODERATOR)

Austin Cory Bart is the lead developer for the CORGIS project (https://think.cs.vt.edu/corgis), which seeks to make a **C**ollection of **R**eally **G**reat and **I**nteresting data**S**ets available to computing educators [2]. The CORGIS site has over 40 datasets in a range of topics including health, literature, geological sciences, history, and much more. Each dataset is available in various data formats, through convenient language bindings, and in a powerful exploratory web interface. In addition to the datasets, the CORGIS project has infrastructure for managing its complex collection and myriad formats. The development of this infrastructure and these datasets have led to the authoring of an on-line text targeted at educational developers on the subject of preparing pedagogical datasets (https://think.cs.vt.edu/pragmatics).

Pedagogical Datasets are distinctive from regular datasets in that they are meant to be embedded in assignments and projects, and therefore can be tied to learning objectives and a learner context. After collection, their structure and information must be suitably scaffolded to ensure that the learners are capable of processing the data. Usability and navigability are crucial in their design, but there are many other design choices to consider. This presentation will briefly highlight the problems, strategies, and experiences that Bart has encountered while designing datasets for the CORGIS project.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGCSE '18, February 21–24, 2018, Baltimore , MD, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5103-4/18/02.
https://doi.org/10.1145/3159450.3159616

3 KALPATHI SUBRAMANIAN

Kalpathi Subramanian is an associate professor of computer science at the University of North Carolina at Charlotte. For the past few years, he has led a team of investigators and developers as part of the BRIDGES project[3]. BRIDGES (**B**ridging **R**eal-world **I**nfrastructure **D**esigned to **G**oal-align, **E**ngage, and **S**timulate) provides a software infrastructure designed to enable the creation of more engaging assignments in data structures and algorithms courses by providing students with a simplified API (Java, C++ supported, with Python API in progress), allowing them to populate their own data structure implementations with live, real-world, and interesting data sets, such as those from social networks, entertainment, scientific, social or cultural data. In addition, the BRIDGES system provides the capability to visualize and share the constructed data structures via a web link, which can be highly engaging to freshmen/sophomore students.

In this talk, Dr. Subramanian will present recent work using BRIDGES, new datasets that have been integrated into BRIDGES, its impact on students over the past few years, and experiences of external users of BRIDGES. A key and current goal of BRIDGES is to continue interfacing with new datasets, creation of highly engaging assignments, and mechanisms for dissemination to the larger CS education community. In this regard, work is underway on building interfaces to CORGIS datasets, led by Bart et al. [2].

4 RUTH ANDERSON

Ruth Anderson is a Senior Lecturer in computer science at the University of Washington. For the past four years she has taught CSE 160 Data Programming [1], which aims to equip students with the skills necessary to use programming to answer questions about data sets. The course is an introduction to programming that teaches the basics of Python with a focus on language constructs needed to process data files. Assignments are drawn from science, engineering, business, and the humanities and expose students to a variety of data types (output from a DNA sequencer, images, text, election results, social networks). For early assignments students are given starter code that handles much of the file reading. Later assignments introduce students to the concept of data cleaning and give them experience doing so. The course culminates in an individual or partner project of students' own design. Students are required to propose a data set they will use and an analysis they will complete. They independently design and implement a program to answer a question about a data set and present their results in written and oral formats. The goal is that by the end of the course students can use programming to pose and answer interesting questions about data sets they care about from a variety of domains.

5 NADEEM ABDUL HAMID

Nadeem Abdul Hamid is an associate professor of computer science at Berry College. For the past several years, he has been developing a software framework [8] to facilitate the use of online data sources in introductory computer science courses. The Sinbad (**S**tructure **IN**ference and **B**inding **A**utomatically to **D**ata) library enables students to easily access online data sources in standard formats (CSV, JSON, XML) with minimal syntactic overhead and no need to deal with low-level issues related to parsing and extracting data. Given a data source URL, the library (currently implemented in Java with a Python version in active development) infers the structure of available data, downloads, caches, parses, and binds the data to programmer-defined data structures and representations. Nadeem has used this library for three years in an introductory programming course for non-majors, using Processing [6]. Students are provided a set of labs to practice accessing data at increasing levels of complexity - from primitive data and simple objects to lists of objects (where the class definitions can be provided by students themselves rather than by the instructor or the Sinbad library). For the final project in the course, students are then asked to find a data source on their own and develop an interactive GUI visualization of the data using Processing. This presentation will highlight the goals and rationale behind the design of the Sinbad project, technical and pedagogical challenges and opportunities, and reflect on the complementary nature of this project to the approaches adopted by the other panel participants.

ACKNOWLEDGMENTS

The work for BRIDGES is supported by the National Science Foundation under Grant No. DUE-1245841, DUE-1726809. The work for CORGIS is supported in part by National Science Foundation grants DUE 1624320, DUE 1444094, and DGE 0822220.

REFERENCES

[1] Ruth E. Anderson, Michael D. Ernst, Robert Ordóñez, Paul Pham, and Ben Tribelhorn. 2015. A Data Programming CS1 Course. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE '15)*. ACM, New York, NY, USA, 150–155. https://doi.org/10.1145/2676723.2677309

[2] Austin Cory Bart, Ryan Whitcomb, Dennis Kafura, Clifford A Shaffer, and Eli Tilevich. 2017. Computing with corgis: Diverse, real-world datasets for introductory computing. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. 57–62.

[3] David Burlinson, Mihai Mehedint, Chris Grafer, Kalpathi Subramanian, Jamie Payton, Paula Goolkasian, Michael Youngblood, and Robert Kosara. 2016. BRIDGES: A System to Enable Creation of Engaging Data Structures Assignments with Real-World Data and Visualizations. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE '16)*. ACM, New York, NY, USA, 18–23. https://doi.org/10.1145/2839509.2844635

[4] Thomas H Davenport and DJ Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to findâĂŤAnd the competition for them is fierce. *Harvard Business Review* (2012), 70.

[5] Peter DePasquale. [n. d.]. Exploiting On-line Data Sources As the Basis of Programming Projects. In *Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '06)*. 283–287.

[6] Processing Foundation. 2017. Processing. http://processing.org/overview/. (August 2017).

[7] Olaf A. Hall-Holt and Kevin R. Sanft. [n. d.]. Statistics-infused Introduction to Computer Science. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE '15)*. 138–143.

[8] Nadeem Abdul Hamid. 2016. A Generic Framework for Engaging Online Data Sources in Introductory Programming Courses. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '16)*. ACM, New York, NY, USA, 136–141. https://doi.org/10.1145/2899415.2899437

[9] Daniel E. Stevenson and Paul J. Wagner. [n. d.]. Developing Real-world Programming Assignments for CS1. In *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITICSE '06)*. 158–162.

[10] David G. Sullivan. [n. d.]. A Data-centric Introduction to Computer Science for Non-majors. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education (SIGCSE '13)*. 71–76.