A. Title Page

# A Comparison of Instructional Data Sources in Computing

Dr. Matt Brown, Department of Computer and Information Science

B. Restatement of problem researched or creative activity

With the storage of information exponentially increasing, instructors in the field of computing are faced with a significant challenge in preparing their students for working with the magnitude and diversity of stored data they will encounter [1]. Instructors must identify data for examples, assignments, and exams that provide the necessary meaningful content. Researchers in the field of computer instruction have discussed the advantages and disadvantages of using simulated data [2], real data [3], and textbook data [4] in the classroom, largely considering each type of data individually. In contrast, this research sought to provide a more straightforward comparison of these three sources of data. This would be achieved through proposal of criteria for quality instructional data and a comparison of the advantages and disadvantages of each type of data.

C. Brief review of the research procedure utilized

This research employed a qualitative longitudinal study of thirty sets of instructional data, from each of the three categories of instructional data, used over eleven years of computing instruction in industry and university settings. The advantages and disadvantages of each set of individual data were described and generalized into nine desirable qualities of instructional data. Those qualities were then used to discuss the advantages and disadvantages of real, textbook, and simulated sources of data for computing instruction.

D. Summary of findings

The following nine qualities for instructional data used in computing curriculum were proposed (described below in no particular order):

1. Authenticity: Data sets are from a business, organization, or real process, or indistinguishable from such real data.

2. Quantity: Data sets include an adequate number of tables, tables include an adequate number of rows and columns, (or lists are of adequate length, etc.).

3. Complexity: Data sets are of appropriate complexity, e.g. tables that require joins in a certain order, or data to model that have a mixture of needed and unneeded variables that may require transformation, etc.

4. Quality: Data sets contain realistic information quality issues.

5. Currentness: Data sets are not on topics that are out of date or no longer of relevance to students.

6. Interestingness: Aside from being authentic and current, data sets spark student involvement.

7. Fit to learning objectives: Data sets facilitate teaching of learning objectives or assess students' comprehension of a learning objective.

8. Adaptability: Data sets are adjustable to instructors needs, easily changed to address student questions, to simplify learning objectives, or to propose more difficult challenges based on in classroom feedback.

9. Availability: Data sets are easy to acquire or create and use for both instructors and students.

Under careful scrutiny with the nine proposed qualities, real data may have fewer advantages in regards to the criteria of authenticity, quantity, complexity, quality, and currentness. Donated real data is likely a small subset when compared to the company's full database, carefully screened so that it has no strategic value that competitors could pilfer, cleaned to remove any significant information quality issues (believing they are doing users a favor), parts of the data pertaining to individuals are disguised for confidentiality, not continually updated as it would be at the company. Thus, through the process of donation, real data may become "nonrealistic".

Simulated data may compare surprising well to real data for the criteria of authenticity, quantity, complexity, quality, and currentness Simulated data can be: patterned after real data (resemble the probabilistic tendencies of the real data), as complex as needed, created to include quality issues, as large as needed, updated in real-time with new data by additional simulation. However, availability may be as big of a practical disadvantage for simulation as it is for real data if the expertise and resources for simulation are not available. Textbook data may hold the biggest advantage over real data and simulated data sets in regards to availability. Table 1 summarizes a comparison of the three sources of data according to the nine criteria.

**Table 1. Summary of Criteria and Data Types that Best Fit**

| Proposed Criteria for Instructional Data Set | Suggested Type of Data That Best Fits Criteria |
|---|---|
| Authenticity | Real Data[1] |
| Quantity | Simulated Data |
| Complexity | Real Data[1] |
| Quality | Real Data[1] |
| Currentness | Simulated Data |
| Interestingness | Simulated Data |
| Fit to learning objectives | Textbook Data[2] |
| Adaptability | Simulated Data |
| Availability | Textbook Data[2] |

[1] If real data are an insignificant sample, have been altered to hide information, or cleaned before donation, simulated data may better meet these three criteria.

[2] If an instructor has objectives in addition to those in a textbook, and the instructor is familiar with tools and techniques for simulation, simulated data may best fit these two criteria.

## E. Conclusions and Recommendations

It is concluded that all three sources of data have a place in computing curriculum because of the distinct advantages offered by each. However, it was also noted that well-simulated data may provide more advantages than real and textbook data for computing instruction in many situations. For this reason, it is recommended simulated data be more widely considered for use in computing instruction.

This research was published in [5], where more details regarding the findings, conclusions, and recommendations may be found.

## REFERENCES

[1] Fayyad, U. & Uthurusamy, R. Evolving data mining into solutions for insights. *Communications of the ACM, 45*(8), 28-31, 2002.

[2] Douglas, D. , Cronan, P., and Jensen, B. Microsoft Enterprise Consortium How To's for the Classroom. *AMCIS 2010 Proceedings* 2010, Paper 528, http://aisel.aisnet.org/amcis2010/528, retrieved November 28, 2011.

[3] Zhou,Y. and Talburt, J. Staging a Realistic Entity Resolution Challenge for Students, *The Journal of Computing Sciences in Colleges*, 26(5), 88-95, 2011.

[4] Jukic, N. and Gray, P. Using real data to invigorate student learning, *SIGCSE Bulletin*, 40(2), 6-18, 2008.

[5] Brown, M. A comparison of instructional data sources in computing: the case for more and better simulated data in the classroom, *The Journal of Computing Sciences in Colleges*, 27(5), 45-51, 2012.

# A COMPARISON OF INSTRUCTIONAL DATA SOURCES IN COMPUTING: THE CASE FOR MORE AND BETTER SIMULATED DATA IN THE CLASSROOM[*]

*Matt Brown*
*Department of Computer and Information Science*
*Arkansas Tech University*
*Russellville, AR 72801*
*479 356-2161*
*hbrown11@atu.edu*

## ABSTRACT

This paper describes and compares three sources of data used for computing instruction: real data, textbook data, and simulated data. Based on a qualitative examination of thirty data sets used in nine different courses, criteria for comparing the three sources are proposed and used to discuss the merits of each type of data. It is suggested that all three sources should be used because each has its distinct advantages. However, the argument is also given that, when appropriately created, simulated data may provide the most advantages for computing instruction. The paper concludes by making recommendations for creating well-simulated data for use in the classroom.

## INTRODUCTION

With the storage of data exponentially increasing, computing students entering the workplace need to be equipped to work with a variety of data. How to best prepare students for the data they will face throughout their career is thus a significant topic within computing instruction. The purpose of this paper is to describe, compare, and make recommendations regarding the different sources of instructional data used in computing courses.

Many courses within computing curriculum require example sets of data. A number of courses are data driven requiring diverse and large amounts of data, e.g. courses on data mining, databases, information quality, etc. Many other courses may be less dependent on data, but can still be enhanced by appropriate example data, e.g. artificial intelligence, operations research, programming, etc. Within these courses a minimum of three sets of distinct data per topic may be needed, one for an example, one for homework, and one for an exam. If students need to see a variety of scenarios for each topic, the demands for data become even greater. Instructors are faced with a significant challenge in acquiring appropriate and meaningful data to support computing curriculum.

Sources for instructional data may be categorized into three principle groups: real data, simulated data, and textbook data. The term *real* data in this paper is used to refer to data sets that are donated by companies or organizations directly to schools or to data repositories for educational or research purposes. Simulated data are generated by the instructor or students. Simulated data may be computer generated or generated by some physical activity. In origin, data found in textbooks may be either simulated or real, but these data have characteristics distinct from simulated and real data and so are considered a separate category in this paper. Textbook data sets include data provided in a textbook, by a publisher's website, author's website, or companion software. In this paper, data sets within help documentation for a particular software or development platform are also considered a textbook source of data. These three categories, believed to cover the majority data used in the classroom, are compared in this paper. Although the case is made that all three have distinct advantages that make all three needed in the classroom, simulated data sets have the key advantage of being entirely under the control of the instructor. Therefore, the paper concludes by providing general recommendations for simulating instructional data sets.

## BACKGROUND

The benefits of real data have been well documented, e.g. [1, 2]. Such benefits include increased student interest, realistic data characteristics such as size, quality, and complexity, and preparation of students for future work environments. Furthermore, real sets of data are becoming more readily available, such as the repository available through the Enterprise Systems program at the Sam M. Walton College of Business [3]. In this repository alone there are four real databases from four different companies, containing over forty tables and over 147 million rows of data.

Despite the benefits, there are also drawbacks of using real data in the classroom. One disadvantage of real data is the lack of availability of such data for the multitude of scenarios that must be covered in teaching [4]. Another is donation bias, where the most interesting and informative data, i.e. those data providing a competitive advantage, would be the least likely to be given to a university. Also, individual privacy concerns prevent some data from being released by organizations [4]. While examples of real data may be invaluable to spark student interest and emphasize issues such as size, complexity, and quality, the disadvantages of real data effectively prevent real data from being the only source of data for instructional use.

Similarly textbook data have both advantages and disadvantages. Data supplied by an author or software developer have the advantage of being tailored to emphasize topics exactly as they are introduced in the text or as needed by the user of software. Furthermore, textbook data are most often the easiest for instructors to use. However, data supplied by a textbook have multiple disadvantages. Textbook data will likely suffer from being of unrealistic size, unrealistic quality and often be negatively perceived by students and instructors. Sentiment regarding using this type of data alone is expressed by [1] stating that textbook data "under no circumstance prepare the students for the true feel and experience of the massive data sets and the database problems they will need to cope with once they graduate and work in the real world".

The advantages and disadvantages of simulated data for teaching are also discussed in literature. In general, computer-based simulations for any use have benefits such availability, the ability to compress time, and the ability to examine a variety of scenarios [5]. These advantages directly carry over to the creation of simulated data for instructional use. Further, there is evidence that simulated data can aid students in the understanding of abstract and difficult concepts [6]. Disadvantages of simulation include a propensity to oversimplify the simulated data, the need to thoroughly understand the process from which the real data are created, and the required resources to perform a good simulation [5]. At the center of the benefits and drawbacks of using simulated data is the involvement of the instructor. In well-simulated data the instructor would be able to closely mimic real data, adapting the data as needed to particular learning objectives. However, the burden of knowing how to conduct a good simulation in a timely manner then falls to the instructor.

## DESIRABLE QUALITIES OF INSTRUCTIONAL DATA

What are the desirable qualities of data used in computing instruction? To answer this question, thirty sets of data used in nine different courses, in both industry training at university computing classes, over the last eleven years were qualitatively evaluated. Included within these sets of data were eight real data sets, eight textbook data sets, fourteen simulated data sets (four physically simulated by students and ten computer simulated). Each data set was evaluated by examining the reasons why that particular set of data was used, what the learning objectives were as they related to that particular set of data, and the advantages/disadvantages of each set of data. The information from these evaluations was then used to propose the following nine qualities for instructional data sets used in computing (described below in no particular order):

1. **Authenticity:** Data sets are from a business, organization, or real process, or indistinguishable from such real data.
2. **Quantity:** Data sets include an adequate number of tables, tables include an adequate number of rows and columns, (or lists are of adequate length, etc.).
3. **Complexity:** Data sets are of appropriate complexity, e.g. tables that require joins in a certain order, or data to model that have a mixture of needed and unneeded variables that may require transformation, etc.
4. **Quality:** Data sets contain realistic information quality issues.

5. **Currentness:** Data sets are not on topics that are out of date or no longer of relevance to students.
6. **Interestingness:** Aside from being authentic and current, data sets spark student involvement.
7. **Fit to learning objectives:** Data sets facilitate teaching of learning objectives or assess students' comprehension of a learning objective.
8. **Adaptability:** Data sets are adjustable to instructors needs, easily changed to address student questions, to simplify learning, or to propose more difficult challenges based on classroom feedback.
9. **Availability:** Data sets are easy to acquire or create and use for both instructors and students.

It is not suggested that these desirable characteristics would or should be present in every set of instructional data. Trade-offs may be consciously made by an instructor, for example to make data more interesting an instructor may simulate an obviously unrealistic dataset about some current event of interest, trading authenticity for interest. Or, to emphasize a particular concept an unrealistically short, simple, and clean data set may be more appropriate. Or, with need for data to support a particular topic on short notice, availability may become the key characteristic. These nine proposed characteristics do, however, provide a framework for comparing textbook, real, and simulated sources of instructional data.

## COMPARING TEXTBOOK, REAL, AND SIMULATED SOURCES OF DATA

When comparing textbook, real, and simulated sources of data for the criteria of authenticity, quantity, complexity, quality, and currentness, it might seem real data would have distinct advantages, however, in practice this may not be the case. Consider a hypothetical example where a company is donating some of its real data. The donated sample is a small subset when compared to the company's full database, is carefully screened so that it has no strategic value that competitors could pilfer, the data are cleaned to remove any significant information quality issues (believing they are doing the university a favor), and parts of the data pertaining to individuals are disguised for confidentiality reasons. In addition, at the company the data are updated continuously, the donated real data set is a one-time gift, not to be continually updated. In such an example, the donated real data may be only slightly better for the first five criteria than textbook data since instructors are likely to have very little control over these details of the donated data.

On the other hand, well-simulated data for the criteria of authenticity, quantity, complexity, quality, and currentness may compare surprisingly well to real and textbook data sources. Consider a hypothetical example where an instructor takes a sample of real data and patterns simulated data after that sample. The simulated data set is made to resemble the probabilistic tendencies of the real data, the size of the real database, and the quality issues within the database, so that only those familiar with the real database would be able to tell the data are simulated and not real. Furthermore, the data can be updated in real-time with new data by additional simulation.

Likewise, simulated data compares well to textbook and real data sources for the criteria of interestingness, fit to learning objectives, and adaptability. Simulated data can be tailored to contain interesting facts or knowledge concerning individuals or strategic advantages-knowledge that would not be donated by an organization. Textbook data may most directly address learning objectives. However, simulated data by an instructor can be tailored to meet any objective, and particularly those not covered by a textbook. Further, once the steps for creating a particular simulated data set are in place, in most cases altering the simulated data to fit different objectives or to meet the needs of a particular class would be relatively easy.

Textbook data may hold the biggest advantage over real data and simulated data sets in regards to availability. Textbook data should be readily available to both students and instructors with the least amount of effort. Availability is a very real concern with limits on instructor preparation time and class time with students. In regards to availability of real data, data repositories on the Internet have made real data easier to acquire than at any point past. However, the available real data set must fit the topic of instruction-if a needed example does not exist in a repository, to expect an instructor to be able quickly find and persuade an organization to donate suitable data in a brief period of time is unrealistic. In contrast, simulated data can theoretically be made to fit to any topic by the instructor. However, performing a good simulation of data requires an understanding of the real processes and organizations that are being mimicked, an understanding of simulation principles such as statistical distributions and models, and an understanding of the software or programming language for performing the simulation needed.

**Table 1. Summary of Criteria and Data Types that Best Fit**

| Proposed Criteria for Instructional Data Set | Suggested Type of Data That Best Fits Criteria |
|---|---|
| Authenticity | Real Data[1] |
| Quantity | Simulated Data |
| Complexity | Real Data[1] |
| Quality | Real Data[1] |
| Currentness | Simulated Data |
| Interestingness | Simulated Data |
| Fit to learning objectives | Textbook Data[2] |
| Adaptability | Simulated Data |
| Availability | Textbook Data[2] |

[1] If real data are an insignificant sample, have been altered to hide information, or cleaned before donation, simulated data may better meet these three criteria.

[2] If an instructor has objectives in addition to those in a textbook, and the instructor is familiar with tools and techniques for simulation, simulated data may best fit these two criteria.

Because each of the three sources of data for teaching have distinct advantages and disadvantages based on which criteria are considered, it should be clear there is a place for all three in the classroom. The ideal data dependent course may then be one that utilizes all three types of data. Under certain conditions, well-simulated data may have the biggest advantage for the majority of the criteria mentioned, but these benefits are tempered by the expertise needed to generate good simulations. The comparisons are summarized in Table 1.

## CONCLUSIONS AND RECOMMENDATIONS

It is suggested that all three sources of data, real, textbook, and simulated, are needed to achieve the goal of preparing students for the data they will face in the workplace. Real, textbook, and simulated data each have distinct advantages and disadvantages for teaching. However, for the majority of the nine proposed criteria for evaluating instructional data, simulated data sets have the potential to outperform the textbook and real data sources because of the instructor's direct control of the simulation. Therefore, it is recommended that instructors who teach data dependent courses and have not used simulation for instructional purposes consider the merits of such an approach.

To fully realize these advantages the instructor must be versed in good simulation practices. While it is not practical to try to present a tutorial on simulation in this paper, a few general comments and recommendations for simulating instructional data are now given, beginning with where to find tools for simulation. Different types of software for simulation are readily available. Spreadsheet software, although not appropriate for every simulation, can be used to quickly generate many suitable sets of data of modest size (through the use of pseudo random number generation from the continuous uniform distribution and its mapping through an inverse cumulative probability distribution). Special purpose programming languages in the domain of mathematics and statistics, such as R, are particular suited for simulation. R is an open-source statistical programming language that can be used to simulate data with succinct programs [7]. In general, the basic building blocks of simulation such as numerous pseudo-random number generation and statistical distribution functions are available in a variety of programming languages/development platforms.

When possible, instructors should base simulation on real data sets. This includes not over-simplifying assumptions about the processes being simulated. For example, it might be tempting to assume that occurrence of certain events could be simulated following a Poisson distribution, when in fact the data are autocorrelated (occur in clusters of events) and a negative binomial or gamma distribution model is much more appropriate as a basis for the simulation. A sample of similar data, a plot, and a check of the average and standard deviation may be all that is necessary to determine how best to simulate the data. Understanding and simulating potential data quality issues may further contribute to the authenticity of the simulation.

Finally, instructors should look for ways to make the simulation interesting for students. Since the simulation is in control of the instructor, simulated data can often be tailored to include examples that may interest students, involve current events, or even involve data the students themselves generate. Turning the simulation over to the student to further engage the student may lead to better learning and have the positive side-effect of imparting some basic knowledge of simulation to the student.

## REFERENCES

[1] Jukic, N. and Gray, P. Using real data to invigorate student learning, *SIGCSE Bulletin*, 40(2), 6-18, 2008.

[2] Singer, J. and Willett, J. Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis, *The American Statistician*, 44(3), 223-230, 1990.

[3] Douglas, D. , Cronan, P., and Jensen, B. Microsoft Enterprise Consortium How To's for the Classroom. *AMCIS 2010 Proceedings* 2010, Paper 528, http://aisel.aisnet.org/amcis2010/528, retrieved November 28, 2011.

[4] Zhou,Y. and Talburt, J. Staging a Realistic Entity Resolution Challenge for Students, *The Journal of Computing Sciences in Colleges*, 26(5), 88-95, 2011.

[5] Banks, J., Carson, J., Nelson, B., and Nicol, D. *Discrete-Event System Simulation*, Prentice Hall Publishing, 2010.

[6] Mills, J.D. Using Computer Simulation Methods to Teach Statistics: A Review of the Literature, *Journal of Statistics Education* 10(1), 2002

[7] Goodman, D. *Notes on R for Stochastic Simulation and Elementary Statistical Inference*, 2011, http://www.esg.montana.edu/R/rnotes.pdf, retrieved November 28, 2011.