Databases! A Web-Based Introduction to the Data Science Techniques of Database Querying and Design

Jennifer Broatch - Arizona State University

Project PIs: Suzanne Dietrich - Arizona State University, Don Goelman- Villanova University

School of Mathematical and Natural Sciences
Arizona State University

Purpose

Want to introduce data science skills to your introductory statistics students of ANY major?

We present a web-based learning tool using visualizations and animations to introduce database topics to a variety of students with no data science or computing background (http://databasesmanymajors.faculty.asu.edu/).
The interactive modules engage student learning in the three major topics:

1) **Introduction to Databases:** relational databases and how they differ from spreadsheets,
2) **Introduction to Querying**: querying of relational databases,
3) **Database Design**: conceptual design of data, which explains how to model data and then map the design to a relational database schema.

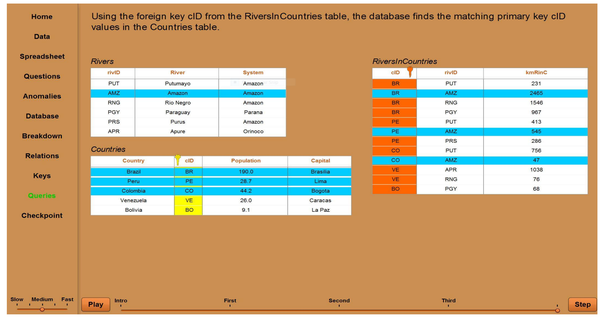
Each module is self-contained and can be completed in about 1 hour outside of class. Each module is presented in multiple STEM domain applications, specifically in statistics, forensics, astronomy and ecology, to attract students of all majors that utilize databases. This early exposure and introduction to database topics has a broad audience and can be used in any introductory statistics, data science, science, or any other course that might want to promote early data science skills.

Benefit

Early exposure to data science skills, such as relational database management skills, are essential for statistics students as well as other disciplines in an increasingly data driven society.[1,3] The concept of relational databases and querying are shown to be essential for data manipulation and management skills.

Introducing database concepts and development through visualization has demonstrated to be a successful pedagogical method for understanding this concept. [2]
After completing the animations, students will have an increased understanding of basic relational databases and querying. The visualization of these operations and how the SQL standard specifies these operations provides a strong foundation for students to start more in-depth coverage of relational algebra and SQL, if desired. Additionally, statistics students can begin to perform statistical analyses of queried results and connect learning to other statistical software packages (SAS and R).

Introduction to Databases

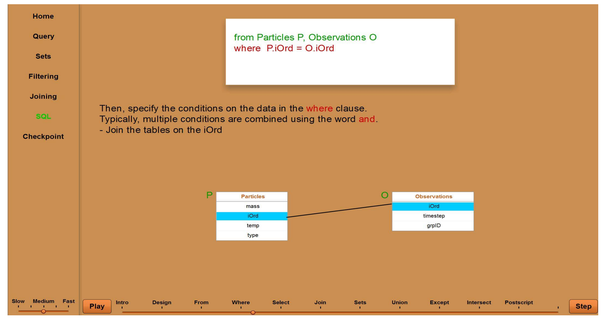


Linking tables with a primary-foreign key (GIS Application)

The **Introduction to Databases** module promotes the understanding of how a database operates, and how to better utilize the power databases have. Databases provide a powerful tool to ask different questions, or queries, of that data without changing the data.

In this module students will learn:
- Limitation of spreadsheets
- Breakdown of spreadsheets into smaller tables to avoid redundancies
- Introduction to primary and foreign keys and how a database uses keys to identify and relate information
- Brief introduction to asking questions over a database

Introduction to Querying

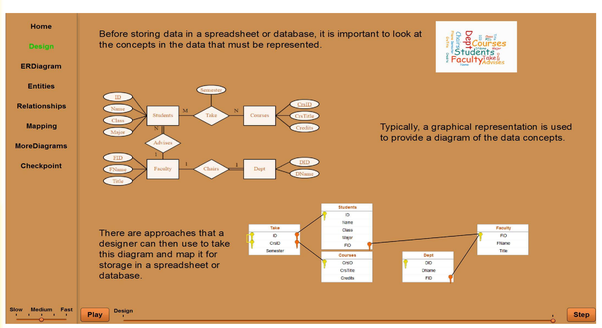


SQL visualization (Astronomy Application)

The **Introduction to Querying** module provides a conceptual introduction to the various operations required to retrieve data from a database to answer a question. The visualization of these operations and how the SQL standard specifies these operations provides a strong foundation for students to use SQL to query relational databases.

In this module students will learn:
- Motivation to identify data and relationships
- Common set operators
- Operations to horizontally and vertically filter data
- More ways of combining tables of data that require a form of filtering
- Introduction to querying using SQL

Database Design



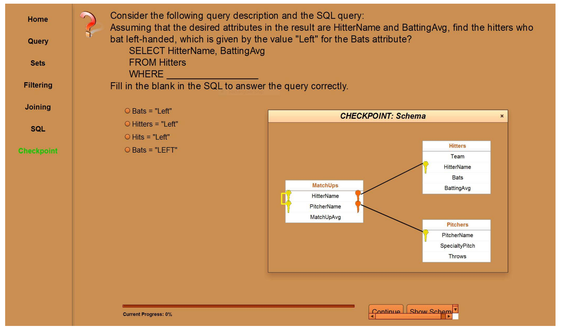
ER Diagram (General Application)

The **Database Design** module builds on the previous two modules with the introduction of entity-relationship diagrams, where entities are associated by the use of relationships. These relationships are realized within a database using referential integrity between the foreign and primary keys, which are essential to the querying process.

In this module students will learn:
- What is conceptual design?
- Overview of Entity Relationship Diagrams (ER Diagram) and how they are mapped to tables
- What concepts are stored in a database and how they are related
- Alternative approaches to ERDiagrams

Formative Feedback

Self assessment quizzes referred to as "Checkpoints" are provided at the end of each module to promote the understanding of the material presented. Feedback is always presented to the student. If a student answers incorrectly, and the question will be asked again later.



Teacher Resources

Resources are provided to teachers to introduce the animations as well as follow up on skills learned using cooperative learning exercises in class, if desired. Students are provided a worksheet for recording their Access query and their formulation of the same query in SQL. The QBE interface of Access inherently supports the design aspect of answering a query and illustrates the primary-foreign key relationships.

To further the connection to statistics, a statistics classroom follow up activity is available that asks students for univariate statistics of queried results.

References

American Statistical Association (ASA) (2014), "Curriculum Guidelines for Undergraduate Programs in Statistical Science," [Online: https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf]

Dietrich, S., Goelman, D., Borror, C. and Crook, S. (2015) "An Animated Introduction to Relational Databases for Many Majors", *IEEE Transactions on Education*, Volume 58, Issue 2, pp. 81-89. [Online: http://dx.doi.org/10.1109/TE.2014.2326834]

Horton, N. J., Baumer, B. S., and Wickham, H. (2015), "Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics," *CHANCE*, 28, 40–50. [Online: http://chance.amstat.org/2015/04/setting-the-stage/]

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DUE-1431848/1431661. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.