NATIONAL ACADEMIES Sciences Engineering Medicine

Foundations of Data Science for Students in Grades K–12 A Workshop

September 13-14, 2022

SEPTEMBER 2022

Real World Application

NASA MY NASA DATA https://mynasadata.larc.nasa.gov

The NASA MY NASA DATA project began in 2004 to make NASA's large collection of Earth science data accessible to K-12 students. The website now offers a number of audience-tested and proven tools for introducing data to students.

Challenge Educators Face

Data science courses are not widely taught at the high school level. Partnerships between colleges, nonprofits, and high schools will need to be established to support, implement, and sustain high-quality professional development, course development. What are examples of successful partnerships to facilitate Data Science teaching and learning at the K-12 levels?

Existing Tool/Software/Technology

Gulf of Maine Research Institute

Our work focuses on explorations of locally relevant impacts of climate change. At its heart, climate change is change over both space and time. Our experience is that existing tools do one or the other well, but not both. So CODAP is fantastic for time series data. Tools from ESRI and Field Scope are great for GIS data. This is complicated by pedagogy. CODAP is designed for a "messing around" approach to working with data, which we love. To our knowledge there's no equivalent that allows play with GIS, particularly over time. Animation of place over time is the likely current solution but I don't THINK (could be wrong!) any existing tools put that power in the hands of learners.

Existing Tool/Software/Technology

CourseKata https://coursekata.org

From their website: "CourseKata Statistics & Data Science
Welcome to CourseKata Statistics and Data Science, an innovative interactive online textbook for teaching introductory statistics and data science in colleges, universities, and high schools.

Part of CourseKata's Better Book Project, we are leveraging research and student data to guide continuous improvement of online learning resources."

Existing Tool/Software/Technology

STATS4STEM stats4stem.org

From their website: "About STATS4STEM

Fully funded by the National Science Foundation, STATS4STEM provides a collection of learning, assessment, tutoring, data, and computing resources for statistics educators and their students. Our assessment and tutoring feature allows for instantaneous student feedback with question specific hints for students who are struggling. In addition, real-time learning charts and reports are available to help educators adapt instruction to meet the individual needs of their students. Our computing resources integrate real-world data with the RStudio statistical computing platform, allowing all students the opportunity to conduct real-world statistical analysis. Finally, in an effort to foster community, our site integrates a message board for statistics educators looking to share ideas, tips, and insights with other educators worldwide.

Any high school, college, or university educator worldwide who teaches statistics or a course related to statistics can register."

Challenge Educators Face

Learning the fundamentals of how data is collected, analyzed, and communicated starts with a solid foundation in the science and engineering practices outlined in the *Framework for Science Education*. One challenge educators face when teaching data science is the lack of fundamentals.

Connecting an abstract concept like data science to real-world applications can be a huge barrier when it comes to keeping students engaged. Many science teachers struggle to engage students actively and feel stuck using slides, lectures, and traditional science experiments to collect data that can be time-consuming to set up and maintain.

Teachers need time to be teachers, and students feel more engaged when they connect data to the world in which they live.

Existing Tool/Software/Technology

Pivot Interactives https://bit.ly/3oL4cWf

Two educators from Minneapolis, Minnesota, created a platform called Pivot Interactives that allows students to explore scientific phenomena and make sense of it through data analysis tools. Their goal is to ignite students' passion for science, and teachers' love for teaching, and to instill the proper fundamentals of collecting and analyzing data.

What makes Pivot Interactives unique?
Pivot Interactives is the only platform for interactive video-based science activities with embedded data-collection tools. Based on real-world science phenomena, students become engaged with data capture, process, analysis, and communication—the fundamentals of data science. Under controlled conditions, students who use a curriculum based on Pivot Interactives interactive video showed significantly more significant learning gains in critical thinking skills even when compared to integrated hands-on learning. Science education researchers at MIT and Carleton College measured students' experience using Pivot

Interactives interactive video compared to traditional methods:

- 85% reported that interactive video made it easier to understand the scenario being investigated
- 92% said they'd encourage their friends to take courses that use interactive video
- 65% wished MIT's online physics course included more of our interactive video
- 60% reported an increase in using scientific skills like measurement techniques when using our interactive video

Pivot Interactives offers a free trial for science educators.

Existing Tool/Software/Technology

Khan Lab School

I piloted a year long data science course to high schoolers (fresh-seniors) last year and it went great! We used the Coursekata curriculum and supplemented with projects and other activities I created. We use Jupyter notebooks along with an integrated textbook and Deepnote.com and R to analyse our data.

Existing Tool/Software/Technology

The FieldScope Project https://www.fieldscope.org

FieldScope is an innovative map-based data collection and analysis platform managed by BSCS Science Learning. It supports citizen and community science projects and other initiatives with tools to collect data from dispersed geographic locations and to analyze trends, patterns and change over time, including the impact of interventions. FieldScope enables organizations, community members and learners to monitor and actively address issues, such as environmental and social challenges, that are concerning to them. It empowers participants in projects to use maps, graphs, and other visualization tools to make meaning of crowd-sourced datasets and turn data into stories that can be used in science communication and grassroots advocacy.

Existing Tool/Software/Technology

Invitations to Inquiry https://bscs.org/resources/invitations-to-inquiry/

Invitations to Inquiry are short instructional activities designed to help middle and high school students work with large data sets hosted on FieldScope. Teachers and students use the interactive FieldScope platform to collect, visualize, and analyze environmental data. With these new Inquiries, students can explore FieldScope's advanced mapping and graphing tools to dig deeper into data in the context of meaningful science classroom lessons. Each lesson engages students in interpreting graphs or maps, or both, and figuring out what the data means. Ultimately, the Inquiries are intended to increase student confidence in working with data and using visualization tools. These Inquiries are designed for 2-4 days of learning and support the Science and Engineering Practices from the Next Generation Science Standards. They include teacher guides, slides, handouts, and other instructional resources and supports.

Existing Tool/Software/Technology

Center for Curriculum Redesign (CCR)

My Center has deeply redesigned Math standards, in partnership with the OECD, to reflect a K-12 emphasis on data science (aka stats/probs) and discrete/computational math. The entire set is visible at https://curriculumredesign.org/modern-mathematics/

Real World Application

Invitations to Inquiry https://bscs.org/resources/invitations-to-inquiry/

The Invitations to Inquiry project is designed to provide short learning experiences (2-3 classroom 45 minute periods) that are focused on using real-world community and citizen science data sets to make sense of a phenomenon. Students are introduced to data sets, and are scaffolded through a process that helps the consider where the data comes from and how it can be used to answer questions (or not). The data is hosted in FieldScope, a citizen science mapping and graphic platform, and the lessons use data from five different citizen science projects. Additionally, we have a self-paced free course for teachers to become familiar with the project, the lessons, and FieldScope. We also provide some program level resources for helping students with data (e.g. using data in discussions, introducing variables, choosing data representations). Importantly, these lessons are intended as introductory learning in data science, and they are explicitly designed to help build teacher- and student confidence in working with data. There are opportunities to go further with data in a few of the lessons (called Data+) activities.

Real World Application

Meharry Data Science Summer Academy

Students, mostly from underrepresented backgrounds, learned to apply coding and data science to robotics on the Meharry Medical College campus. The students were attending the Meharry Data Science Summer Academy, funded by NASA through the Minority University Research and Education project and provided to students at no cost. NASA's goal is to support the dreams of students from traditionally underrepresented and underserved communities to enter careers in science, technology, engineering and math. In addition to courses in programming, robotics and data science, special sessions featured Black professionals in data science and robotics, Kenneth Harris, the deputy lead integration engineer for the NASA James Webb Telescope's Integrated Science Instrument Module Electronic Components, and Dr. Sian Proctor, the first Black woman to pilot a spacecraft. ***The overwhelming message was clear. A future in science and technology is within each student's grasp.***

Real World Application

Databot https://databot.us.com/ds4e-iaq/

Hi, databot™, as the name might indicate, is a company that is pretty passionate about K12 Data Science. We sell a STEM device, databot™, a tiny, wireless, classroom-tough multi-sensor product with 15 built in sensors. Students and educators can easily connect to databot™ via bluetooth and immediately have access to scientific data from databot™ including CO2, VOCs, humidity, air pressure, and many more. databot™ joined the DS4e coalition this summer and pledged to create 10 lesson plans that emphasize data science practices. The first is an air quality investigation and was published a few weeks ago. Please review and share as an example. Feedback is welcome as we work to produce more of these - we'd love to make them better if you have ideas.

Challenge Educators Face

Gulf of Maine Research Institute

Inquiry-, interest-, and place-based pedagogies typically are described as highly engaging for learners whether in the classroom or in an informal learning context like 4-H. Yet these approaches pose challenges to learners having a data-rich experience. Data relevant to explorations of my place or my question can be difficult to secure, to complicated for the developmental stage of the learner, or simply may not exist. Where it exists, data provenance may not be obvious and may therefore hide important information from the novice data scientist. This produces what we call “data dead ends” and frustration – independent learners design interesting questions to pursue but become exhausted at the task of finding and accessing data that might inform their investigations. Where data exists and is usable, many tools don’t support authentic work with the data by novices.

Challenge Educators Face

Gulf of Maine Research Institute

In response to dramatic media reports about massive die-offs of moose, a class in Aroostook ME decides to investigate the impact of ticks on Maine moose populations and its link to climate change. They need to secure and visualize state data on both tick and moose populations alongside precipitation and temperature data over a long enough period to distinguish a climate-driven trend from an anomalous year/period or natural variability. Tick populations also turn out to be related to several key invasive plant species (e.g., barberry) that are tick “magnets.” Ideally they would do so in a tool that allows them to mess around with the data in an inquiry based fashion. And, like most climate phenomena, they need a tool that allows them to view change over both space and time.

Existing Tool/Software/Technology

Data Nuggets https://datanuggets.org

Data Nuggets are free classroom activities that bring authentic data and science stories into the classroom. Over 100 Data Nuggets are available on our website. Written by the scientists behind the research, each activity guides students through the process of working with a dataset to answer a scientific question. Because the authenticity of the research process is maintained, students often face unexpected results, messy data, and findings that do not support original hypotheses. Students who use Data Nuggets show increased interest in STEM careers, confidence when working with data, and improved in their abilities to construct scientific explanations. For updates and announcements when new Data Nuggets are added, follow us @Data_Nuggets!

Existing Tool/Software/Technology

DataClassroom https://about.dataclassroom.com/

DataClassroom U https://u.dataclassroom.com/

DataClassroom and DataClassroom U are web-apps for graphing, statistics, and data analysis in grade 6-12 and university level science and math classrooms. The tool runs on any device that can access the internet and can integrate with learning management systems such as Schoology, Clever, Google Classroom, Canvas, and Classlink The tool has been designed by teachers to provide the opportunity to integrate next-generation data skills seamlessly with the learning experiences they are already creating.

Functionally, students can produce a wide variety of data visualizations, and get support with graph choice, applying inferential statistics, and learning the math behind statistical tests with animation. The DataClassroom U version of the tool contains a Bridge to R which generates well annotated R code for any of the graphs or statistical tests called through the easy to use interface. This supports students early in the process of learning a statistical programming language.

The app contains 100+ curated datasets and lessons that are available with the free version of the app.

Challenge Educators Face

Regardless of how hard we work at designing clear, important, and meaningful District and State science and math standards and guidelines that incorporate data science skill development, what actually ends up happening in the classroom may be vastly different than our intentions. The amount of training on data science that K-12 teachers have and their confidence with teaching it has immense variability, even within a single school, mine included. Many teachers thus either avoid doing much meaningful data science with their students or they transfer their own misunderstandings onto their students by doing it wrong.

Existing Tool/Software/Technology

DataClassroom https://about.dataclassroom.com/

I subscribe to DataClassroom for all of my students. A major focus of DataClassroom is data management and visualization, but the app is also designed as a teaching tool and there are some wonderful interactives for students to see how their data are actually being processed and analyzed. The app is very user friendly and is easy to use.

Real World Application

Making Sense of Data: A Statistics Survival Guide

I have written a statistics guide for my high school students. I call it Making Sense of Data: A Statistics Survival Guide. Here is a link: https://drive.google.com/file/d/18EfKKZrYJdSnsMdhdzwOXWnBuw4Ivbom/view?usp=sharing

Real World Application

The Relationship Between Bone Density & Age in Older Females Anatomy and Physiology Data-based Question (Part 1)

Strong bones are important to human health. Bone mineral density (BMD) is a way to measure how much calcium and other types of minerals are in an area of bone. A BMD test helps health care providers detect bone diseases like osteoporosis and predict a person's risk for bone fractures. Osteoporosis is defined by the World Health Organization (WHO) as a "progressive systemic skeletal disease characterized by low bone mass and microarchitectural deterioration of bone tissue, with a consequent increase in bone fragility and susceptibility to fracture." For example, the figure below compares a healthy femur to one with osteoporosis and a resulting fracture in the neck of the femur. Osteoporosis affects people of all biological sexes and genetic ancestry and can occur at any age, although the risk for developing the disease increases as people get older. For many biological females, the disease typically can begin to develop a year or two before menopause.

Link to the full activity, including the data and citation:
https://docs.google.com/document/d/18MO6avz6YrguPUO8MY8uPLU-0xLpFmiFAwSm1pn77yI/edit?usp=sharing

Real World Application

The Relationship Between Bone Density & Age in Older Females Anatomy and Physiology Data-based Question (Part 2)

Medical researchers at the University of California in San Francisco were interested in tracking just how quickly bone density declines on average in females and how this change might vary in different bones. This kind of research is called a descriptive study. The researchers obtained (with permission) 1,886 bone density scans and age data of several hundred patients from three clinics. The subjects were stratified into five-year age groups and BMD (g/cm2) was calculated for several skeletal areas of interest, including the lumbar spine (L1 - L4), proximal radius, distal radius, and calcaneus. The researchers then converted BMD to a BMD percent relative to the 66-70 year age group. These data are presented in Table 1 on the next page.

1. Create a Google Doc that you will submit to the Google Classroom assignment.
2. Write what you think the original research question might have been for this study.
3. Transfer these data to a Google Spreadsheet. Hint: In your spreadsheet, you will need to arrange the data in a Tidy Data Table with Age grouping, Bone region, and Relative BMD (%) as your three variable column headings.
4. “Fetch” your spreadsheet into DataClassroom.
5. Create a single data visualization that shows how the mean relative bone mineral density (%) changes with female age for each of these three skeletal regions. Hint: When you create your graph in DataClassroom, include Bone region as a Z variable. What should the X and Y variable be?
6. Copy and paste your graph into your Google Doc.
7. Write a brief summary that answers the research question. What can you conclude from your data visualization? Make sure to include data from the graph in your summary.

Real World Application

Indirect Estimate of Potato Cell Cytoplasm Molarity (Part 1)

Background: In 1940, botanists at The Ohio State University, Bernard Meyer and Atwell Wallace, tested two methods for indirectly determining the molarity (moles cytoplasm particles per liter of cytosol) of the solution within potato (Solanum tuberosum) cells. In general, the botanists soaked cores of potato tissue in a series of different sucrose solution strengths. The solutions were measured in osmotic pressure (atmospheres at 20° C) and ranged from 0.0 atm to 33 atm. In their study, a change of zero in length or total weight of the cores indicated that the solution in which the cores were soaked was isotonic to the potato cell cytoplasm. Two figures from their methods investigation are pictured below. [Figures appear in the full activity available in the linked document]

Study the two figures and determine the osmotic pressure at which the solution and cytoplasm are isotonic.

Convert the osmotic pressure in atmospheres to molarity using the equation: [The equations above do not copy and paste well into a Google Form. Equations appear in the full activity available in the linked document.]

You now have an experimental value for the molarity of potato cell cytoplasm.
Procedure: You will replicate Meyer's and Wallace's methods to test their experimental value against your own experimental value. However, you will use % change in mass of potato cores as your final dependent variable and sucrose solutions of 0.0, 0.1, 0.2, 0.3, and 0.4 M as your independent variable. By using % change in mass you will not need to worry much about controlling your initial masses of your potato cores.

Real World Application

Indirect Estimate of Potato Cell Cytoplasm Molarity (Part 2)

[Below I have left out the procedures to hit the character limit]

Calculate the % change in mass of each fry at each solution molarity and record your data in the shared class Google Spreadsheet.

Calculate the class mean % change in mass for each solution, and graph the data with standard error of the mean (SEM) as error bars.

Use these data to determine indirectly an experimental value of the molarity (moles cytoplasm particles per liter of cytosol) of the solution within potato tissue. Illustrate this method visually on your graph.

Perform two statistical tests. One test should show the strength of the relationship and whether or not the relationship is statistically significant. The other test should allow you to determine whether or not at least one of the mean % changes in mass is significantly different than any of the others. In other words, determine whether or not the solution molarity (the independent variable) had a statistically significant effect on the mean % change in mass of the potato tissue (the dependent/response variable). Refer to your Making Sense of Data: A Statistics Survival Guide for guidance.

Challenge Educators Face

Many educators don't receive great training on how to implement a Data Science curriculum, so they fall back on old practices and rely on their knowledge of Statistics to push them through. This usually leads to students being in courses labeled as Data Science, but taught as Statistics, which is to say they omit the computer science aspects that are so necessary to the implementation of a Data Science foundation. To this point, many math educators don't think they belong in the Computer Science field (many also don't believe they belong in the Statistics field, strangely). However in California, you need a math credential to teach Computer Science, so the disconnect is probably as a result of their pre-service courses (many of which lack a Computer Science strand).

Real World Application

Timberline students in Intro to Data Science used data from our school to help determine if students who went from Algebra 2 to AP Calculus performed similarly to students who went from Precalculus to AP Calculus. Their results were used to help the school develop math pathways for all students at Timberline High School.

Existing Tool/Software/Technology

I would like to share an example of how a data tool can be designed and adapted for learners to enable students to better engage in Analyzing and interpreting solar data and phenomena. A Solar Data Viewer from NASA, the Helioviewer https://helioviewer.org/, was redesigned (https://student.helioviewer.org/) to put phenomena exploration first. Designing tools that better enable sensemaking and student exploration will be critical to the future of data science for learning.

Real World Application

Coding Like a Data Miner

We've developed an NSF funded curriculum (Coding Like a Data Miner) that teachers students how to scrape (i.e., data mine) Twitter along their personal interests. The curriculum combines quality computing practices and data science techniques using equity-driven framings. This means rather than navigate data sources generated by others, learners can construct their own data sources in real time. In this way, Coding like a Data Miner is couched in real work applications suited to reflect the nature of the increasingly digital world around us!

Read more about the project here:
https://www.nsf.gov/awardsearch/showAward?AWD_ID=2137708&HistoricalAwards=false

Existing Tool/Software/Technology

Google Map Data: a source of "measurement of the distance" to Students Traveled Paths for supplementing Data Science learning Activity (Part 1)

The growing interest in preparing better K–12 students to work with data (Lee et al., 2021; Philip et al., 2016; Philip et al., 2013; Wilkerson & Laina, 2018; Wilkerson & Polman, 2020) has at its core an urgent need for publicly and readily available data sources (Wilkerson & Laina, 2018). I provide an opportunity for this by building on Philip and Garcia's (2013) assertion that students' experiences outside of the classroom, encapsulated by images, video, sound, notes, and GPS tags, can ever more easily become texts for study (p. 311). Students picking distance measurements from Google Maps, I engage them to locate two locations they have manually walked to represent the path on graph paper/board as a relationship between distance traveled and time used.

During the walk
Students taking a walk aim to pick six situations on their way but not distance. Students note the situation at the beginning and end of their 3 five minute walks, which adds to a fifteen-minute walk and six different positions. The first five minutes involve students walking linear their way forward. Then they turn to their right to walk for the next five minutes. The student then turns to their left for the last five minutes.

After the Walk
Using a graph sheet/board, students will plot a traveled graph in a relationship of distance walked against time used. Students will use Google Maps to pick the respective measure of the distance they travel within the specific spots in the full path to represent a relatively accurate travel path.

Existing Tool/Software/Technology

Google Map Data: a source of "measurement of the distance" to Students Traveled Paths for supplementing Data Science learning Activity (Part 2)

Implication for Data Science Learning
This way, by centering existing data sets in the form of Google Map data as objects of inquiry to students' traveled routes, I positioned students as non-reactive to those data sets but as designers and authors in their rights(Lee et al., 2021).

I show an example of an existing software/tool/technology that teachers can use and could adopt to supplement Data Science learning. Thus, Students who engage in such educational data mining (picking distance from the Google map) tend to focus on using Google Maps as a tool for discovering data. This engages students to build their data and computational literacies.

What additional research is needed?
I propose the activity provides opportunities for using the cross-disciplinary three-part framework-personalized, cultural and socio-political layers that shape who and what is counted, measured, and recorded in data sets (Lee et al., 2021).

A call for research on Data (Science Learning) as our Gateway to nearing True integrated STEAM. This is because others like "STEM +C" are mostly not focused on the integrated nature of STEAM but its part, Computer Science.

Existing Tool/Software/Technology

For the past three years, I have been working on transitioning from a traditional high school Statistics course to a one that uses real-world applications and daily work with SAS Studio. Thus Statistics at Perry High School has been replaced with an Intro to Data Science course.

I have been very fortunate to be supported by Professor Lisa Dierker from Wesleyan University. Professor Dierker is a co-creator of Wesleyan's "Passion-Driven Statistics" model, a data-driven, project-based introductory curriculum backed by the National Science Foundation. This flexible curriculum engages students from a range of disciplines with large, real-world data sets and code-based analytic software(SAS Studio), providing experience in the rich, complicated, decision-making process of real statistical inquiry. On a daily basis, students work with current real world data sets such as Gapminder, Nhanes, Addhealth, Nesarc… I found the Passion-Driven Statistics model really enriched my teaching materials and student engagement.

Existing Tool/Software/Technology

This year I am involved in the Ohio Data Science Foundations Course Pilot: using the Data Science Curriculum(UCLA: IDS). By participating in this pilot, I will be able to bring these two models together and finalize a course that is aligned with the state standards and provide my students with real-world engaging learning experience.

Also this year, two students from the Intro to Data Science from last semester, are now enrolled in our Apprenticeship class. The plan is they will prepare for the SAS certification test and explore options with local companies. Our school has just been approved to access the SAS Educator Portal; all Perry Lake students will be able to access the SAS Skill Builder for Students. These students will independently work through the SAS Virtual Learning Environment: SAS Programming 1: Essentials course to prepare them for the SAS certification. I believe these students will be the first in the state to add this credicential to their resumes.

Real World Application

At the beginning of the course, the DSS classes collect data about themselves (age, height, shoe size, cotton ball toss, reaction speed, and more) to analyze throughout the course. The data collected was also geared to have a mix of quantitative and categorical data to use for different group and regression models analysis. When analyzing the data, we are able the relate topics of sampling, measurement error/bias, mistakes, and outliers to real data they collected. I believe this is more powerful because it is one thing to be given a data set and talk about measurement bias/mistakes in the data but it is another when they are the source of the data. One example of a good discussion was "How did someone get a measurement of 8 inches for their index finger?"

Challenge Educators Face

My colleagues and I led a state-wide professional development course on "What is Data Science" for High School MATH teachers last year. During this course the teachers identified challenges they face when working to incorporate Data Science activities. Some of these challenges included: time restraints - including the excessive number of math standards they are expected to cover; scheduling restraints - many teachers in Idaho teach in rural schools where there is often only 1-2 high school math teachers for the entire school; to add in a full data science class doesn't seem feasible to them; their own fear or lack of time to learn a new software/tool/technology in order to teach data science - we are thinking as a state to offer a follow up PD course that is intentionally designed to help 6-12 educators get over this hurdle; explicit connections between Data Science and Mathematics Content Standards - many MATH teachers want to know they are still teaching math when they are leading Data Science Activities.

Existing Tool/Software/Technology

Schoolytics

Schoolytics is a student data platform that gives data superpowers to teachers, administrators, parents, and students by building and automating data pipelines, and providing out-of-the-box data visualizations of academic data. We believe in empowering students with actionable data on their own progress and contend that students can and should have agency over their own learning. Schoolytics transforms Google Classroom assignment data into meaningful KPIs and time series trends for users to reflect on and respond to.

By using their Schoolytics dashboard, students get exposure to simple statistics and graphs based on their own data, which takes on a personal and real quality that more theoretical discussions on data science struggle to match. In addition, because the primary source of data for Schoolytics is Google Classroom, teachers and students can engage together and work as partners in analyzing and interpreting the data.

Challenge Educators Face

Best wishes. In science education research, students usually analyze data using the Excel program, because of the advantages of this program students can detect and design the data analysis formulas used. However, when using SPSS, students can only output but do not know how to process data analysis in SPSS. This is an obstacle that is often faced by students when the results of the analysis are presented, where they are not able to explain how the data analysis process is carried out.

As an answer to these problems, I often suggest to students that before using SPSS, first read the SPSS Manual to get an in-depth picture of the contents of the SPSS program, especially related to the use of statistical formulas in SPSS.

Real World Application

Mobile City Science (Part 1)

Mobile City Science (MCS) is a series of design studies of youth learning how to use data about their daily lives to make evidence-based recommendations for community change. These efforts required sometimes large groups of community stakeholders to support youth in this endeavor, including public school teachers and administrators, out-of-school time educators from libraries and museums, caregivers and families, Mayor's Office representatives, and researchers. We developed a broad design commitment to honor and elevate data collected by youth while supporting them to use tools and create data stories in forms familiar to decision makers. This design commitment came from observations and analyses of interactions between residents and urban planners that were talking, and making decisions about the future built environment, using complex spatial data visualizations.

In a Queens (New York City) version of MCS, groups of youth told different data stories in their final recommendations for community change. Some used spatial data and maps at different stages of the MCS experience while others focused almost solely on how and what they saw and felt when moving through the neighborhood. Other groups incorporated major themes from interview and archival data. The case I focus on for this purpose is of two high school students, Jae and Ana, working across various data "kinds" in their creation of a bus tour. They used extant spatial data of the area to supplement data they collected of their own corporeal mobility. Synthesizing these various data kinds, they created a data story, as a layer on a map, representing a future-looking, possible pathway important for community pride and sustainability.

Real World Application

Mobile City Science (Part 2)

Jae and Ana were surprised to discover, via an internet search and interviews of neighbors, that Queens had no tourist buses that showcase the racial diversity and cultural assets of the borough. They found a map comparing the demographics of all the New York City boroughs and found that Queens is the most racially diverse. In showing this map during a community event –an audience of approximately thirty educators, peers, researchers, and community stakeholders (Jae and Ana stood at the front of the room next to a screen displaying their Google Slides)--Jae exclaimed while pointing to a racial dot map, “Look at our beautiful Queens. It’s so colorful! There are White people, there are also a lot of South American, a lot of Asian people, and there is a lot of diversity in Queens.” From this orienting perspective, Jae and Ana continued to build an argument around tourists getting a comparatively white-washed version of New York City when they only visit Manhattan via available tours.

Their solution was a new bus tour so people, as Ana said, “can visualize the places that are really pretty, places that have different meanings to it and attract tourists to come to Queens.”

Existing Tool/Software/Technology

University of Virginia School of Data Science

I have developed a high-school level data science course with one of my students. The course launches this week at a private school in Charlottesville, VA. The course does two things:

1) Students work together in small teams on an end-to-end data science project, which covers the Big Idea of a data science pipeline. Students select a project from an open source repository of municipally-generated data.

2) Students learn the essential ingredients to accomplish (1); they learn a mix of computing, data skills, modeling and statistics.

It would really be wonderful to make this course broadly available so that others can use it.

Real World Application

NSTA Daily Do Playlist

This is an NSTA Daily Do Playlist of three lessons forming an instructional sequence in which students are using mathematical models (computer simulations) to explain the phenomenon of the spread of COVID19. Students use data science and new understandings to make recommendations to keep themselves, their families and their communities safe.

Existing Tool/Software/Technology

Awash in Data https://www.concord.org/awashindata

Awash in Data is an always-evolving e-book by Tim Erickson introducing students (and teachers) to basic ideas in data science. It begins with a set of lessons that take about three hours of class time, plus homework and a mini-project. Students use CODAP throughout. The book itself is "live" in the sense that you can do many tasks in the book itself, that is, CODAP windows are embedded in the website. Like CODAP, the book is free and requires no sign-in to use. The introductory lessons have been used several times with high-school students as a supplement to an Applied Math class. Erickson and Chen published an article describing the work. [Erickson, Tim and Chen, Ernest. 2021. “Introducing data science with data moves and CODAP.” Teaching Statistics. https://onlinelibrary.wiley.com/doi/10.1111/test.12240.]

In 2020, Erickson used the book as the foundation for a one-semester high school introduction to data science. He is updating the book with assignments and learnings from that longer class.

Challenge Educators Face

With regard to teaching science, it is very challenging to find relatively clean, multivariate datasets that allow for rich exploration of science topics that are directly relevant to what teachers are currently teaching. That is, in order to incorporate data science into an already over-full, year long curriculum, teachers need to access multivariate datasets that will enable their students to learn the science by working with the data. Such datasets are hard to find, and take a lot of digging to find—even when they exist.

Real World Application

Fall Data Challenge

Fall Data Challenge sponsored by the American Statistical Association. Each year, this contest challenges undergraduate and high school students to work in teams to analyze real-world data and make recommendations to combat critical issues.

Real World Application

Census at School

Census at School - U.S. is a free international classroom project that engages students in grades 4–12 in statistical problem solving using their own real data. Students complete an online survey, analyze their class census results, and compare their class with random samples of students in the United States and abroad.

Real World Application

ASA Data Visualization Poster Competition

The ASA Data Visualization Poster Competition is for grade K–12 students to create a display containing two or more related graphics that summarize a set of data, look at the data from different points of view, and answer specific questions about the data.

Existing Tool/Software/Technology

GAISE

Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework provides recommendations and a curriculum framework with examples for teaching statistics in the pre-K–12 years.

I am putting GAISE in the “tool” category because learning needs to be assessed and the document provides direction for assessment. There are also many useful examples help teachers see the path to integrating a data-centric approach while satisfying existing curricular demands.

Existing Tool/Software/Technology

When educators are ready to develop their skills in teaching data science and statistics, they can use online professional learning courses and platforms to engage with highly effective activities, videos, data investigations, readings, and assessments. Since 2015, 6,500+ educators from over 90 countries have engaged in such self-guided learning through the Friday Institute for Educational Innovation at NC State University.

Two current self-paced professional learning opportunities include:
Amplifying Statistics and Data Science in Classrooms http://go.ncsu.edu/amplifystats

and

InSTEP: Invigorating Statistics and Data Science Teaching Through Professional Learning
http://instepwithdata.org

Challenge Educators Face

Judges in K-12 science & engineering fairs often find students have very poor grasps of basic statistics and probability, and nearly zero comprehension of error sources and variability. We cannot leave statistics to a HS AP statistics course! Data science basics need to include the basics of these important mathematical concepts, and teachers need support to do that inclusion. More data is NOT equal to better data, necessarily; evaluating data quality is similar in concept to evaluating bibliographic sources.

Challenge Educators Face

Enhancing Statistics Teacher Education through E-Modules [ESTEEM] (NSF DUE 1625713) http://go.ncsu.edu/esteem

The ESTEEM project began in 2016 to develop teacher education curriculum materials designed to support secondary (grades 6-12) mathematics teachers to learn to teach statistics. The project's focus on the statistical education of teachers was due to increased expectations for students to learn statistics at the secondary level that needed to be matched by enhancement and prioritization of statistics teacher education. Lovett and Lee (2017) found that secondary preservice mathematics teachers were leaving teacher preparation programs feeling least prepared to teach statistics out of all content strands they may be responsible for teaching. Hence, creation of high quality, modern statistics (and data science) teacher education curriculum materials was identified as important for the field of mathematics teacher education. Mathematics teacher preparation programs vary widely, and statistical content and pedagogy may be introduced in different courses including a general mathematics methods course, a course on teaching and learning statistics, a statistics content course, or courses focused on technology for teaching mathematics. Course modalities also vary greatly across programs, and there is an increased need for resources that support online learning. The ESTEEM project packages materials into e-modules. Their modular format makes their use more flexible for faculty, allowing them to choose the modules that work best in their teacher preparation program. The modules are easily imported into Learning Management Systems [LMSs] for adaptation and integration with other course materials. Faculty can access the entire set of ESTEEM-developed modules at the ESTEEM portal, available through free registration. At the ESTEEM portal, faculty can download a version of the complete set of materials in an easy-to-use format that can be imported into CANVAS, Moodle, or Blackboard. We also offer a Common Cartridge format that can be imported into other LMSs. All materials are distributed using the Creative Commons Attribution Noncommercial Share-Alike 4.0 license.

Real World Application

Hub for Innovation and Research in Statistics Education [HI-RiSE] (Part 1)

For many years, the projects within the Hub for Innovation and Research in Statistics Education [HI-RiSE] at NC State have worked in classrooms with students engaging them in data science activities using larger multivariate datasets in CODAP. These videos represent a sample of the successful ways students engage in data science skills and thinking within mathematics curriculum and classrooms. We use these videos to support teacher professional learning in statistics and data science.

Video 1: Students Working on the Roller Coaster Investigation
https://www.youtube.com/watch?v=RvzAxKlHr0E
Brief description: A nine-minute video illustrates brief episodes of students in 6th grade, 7th grade, and high school AP Statistics as they investigate roller coasters with multivariate data in CODAP. Consider how the different groups of students utilized statistical and data habits of mind, including the role that context played in their discussion; the ways they engaged in posing questions of interest; the ways CODAP supported or hindered the students' work; and the opportunities for each student to express their understandings.

Real World Application

Hub for Innovation and Research in Statistics Education [HI-RiSE] (Part 2)

Video 2: Discussion of the Roller Coaster Investigation
https://www.youtube.com/watch?v=ETNF_542DvU
Brief description: An eleven-minute video shows a teacher implementing the Five Practices For Orchestrating Productive Discourse during a 7th grade statistics lesson, including anticipating their responses, monitoring and supporting students as they work, selecting and sequencing students' work to share, and facilitating discourse by connecting and building on students' thinking. Reflect on how the teachers' interactions with pairs of students moved their reasoning forward, why the teacher chose the sequence of students' sharing their work publically, and how she used students' ideas to build connections between the student thinking shared.

Video 3: Investigating Fuel Efficiency in AP Statistics
https://youtu.be/HqHiFrI6i-E
Watch a video of high school AP Statistics students using random samples of multivariate data about vehicles to examine relationships between variables. This is an example of how having a real world context and a large sample of data with many variables affords opportunities for students to investigate real issues and learn important statistical concepts.

Challenge Educators Face

Data science is interdisciplinary in nature. However, the core curriculum places most of the work in developing students' understanding of data within mathematics. In taking this approach, mathematical techniques tend to get foregrounded, and other aspects of critical data literacy remain underdeveloped. To better prepare students as citizens in an increasingly data driven world, we need students to have opportunities to engage with data across the curriculum. This raises the challenge of better understanding how data science education intersects with subjects matter taught across the curriculum, the pedagogical content knowledge teachers from different disciplines draw upon as they integrate data investigations with existing curricula, and the resources teachers need to do this.

Real World Application

In the Spring of our High School Intro to Data Science class, students create their own research project. The create their own research questions and create surveys. They collect data, clean it, and analyze it using R Studio and other tools. They even use Code.org to use Machine Learning to create an app based on their data that will make predict/classify new users. They then present their research. See an example below of students who were interested in health in our school community.

https://docs.google.com/presentation/d/17rc7a1JPyU7kGSGWPrYEbXYsEIGP4sh4S5scc2SsAQc/edit?usp=sharing

Existing Tool/Software/Technology

GAISE II (Part 1)

In 2015, the American Statistical Association (ASA) published the Statistical Education of Teachers (SET) report. The report was summoned to further unpack the recommendations of the Mathematical Education of Teachers II (MET II) report, which specified that mathematics teachers especially need preparation in statistics. In 2020, the Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education report was co-published by the ASA and the National Council of Teachers of Mathematics (NCTM). GAISE II incorporates enhancements and new skills needed for making sense of data today while maintaining the spirit of the original Pre-K–12 GAISE report published in 2005.

Now more than ever, it is essential that all students leave secondary school prepared to live and work in a data-driven world, and the GAISE II report outlines how to achieve this goal. To reach the goal of a data-literate population, teachers must be prepared to deliver statistics and data-science content in the classroom.

Existing Tool/Software/Technology

GAISE II (Part 2)

Access to resources that addresses current school-level standards and recommendations put forth in the SET report would empower teachers and teacher educators to teach statistics in a way that is rich and relevant. The book, Statistics and Data Science for Teachers, aims to provide teachers with a foundation in statistics and data as outlined by SET and content of GAISE II included in state standards. In the spirit of GAISE II, this book presents statistical ideas through investigations and engagement with the statistical problem-solving process of formulating statistical investigative questions, collecting/considering data, analyzing data, and interpreting results. For each investigation, worksheets prepared by teachers to be used in the classroom can be downloaded.

This book encompasses all grade bands of teacher preparation (elementary, middle, and high) up to the content of an AP statistics course. The authors envision that it could be used to guide entire courses and professional development, or portions of courses and professional development that teachers may be taking. A main goal of the book is to provide teacher educators with a resource to use when preparing teachers of all grade levels to teach statistics in their classrooms. The goal of this book is to provide guidance in preparing educators in a way that helps teachers gain:

- an understanding of statistics and data-science content covered in grades K–12,
- an appreciation of and a familiarity with using technology such as apps and statistical software, and
- an understanding of how to think about data.

Challenge Educators Face

Data Science is so new that students do not have a fundamental understanding of what it is, or even who uses it, which makes diving into curriculum difficult. Students also have a false sense of the education that is necessary to pursue a career in a data field because they are under the false assumption that anything that is in math must require a college education and a great deal of higher math. All of this combined scares students and creates a hesitation around the field. To alleviate this roadblock this year, I began my year with two activities during the first week of school. The first thing I did was have students look at fundamental data literacy skills. Students worked in groups to learn about a skill and made a mini-poster explaining the skill in their own words and shared out with their classmates. Students had an immediate introduction to what data science and data literacy is though this activity and were able to easier transition into the curriculum with a basic background of what this class would cover. The second thing I did was a career project. Student pairs chose a career to research that uses data science and were given a checklist of basic information to find. Then, they were tasked to create a one-slide presentation showcasing their career. Students then presented their findings to their peers. The results were very positive. Students realized from the first week of school that there are careers they never heard of that are in high demand right now. Many saw that some jobs require a certificate, or Associate's degree, to be qualified, rather than a Bachelor's or Master's, and others also found that many companies will pay their employees to obtain higher education once employed and moving up. All of this, along with the average salaries presented for each career, students became excited about the possibility of opening up new future opportunities.

Real World Application

Using CODAP to Tell Different Stories is an online lesson in The Statistics Teacher, a joint online publication of the American Statistical Association and the National Council of Teachers of Mathematics. This lesson uses the Statistical Investigation Cycle at the middle grades level to engage students in using and developing their understanding of statistics through problem solving with technological tools. Students collect data, analyze and interpret the results, as well as review the work of their peers to more deeply understand the meaning of the data and its potential implications. The problem solving and sense making using the various facets of data science provides an exceptional learning experience for students.

Real World Application

The Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) document, from the American Statistical Association and the National Council of Teachers of Mathematics, provides a wealth of real-world applications organized around consideration of the development of statistical thinking/data science in PreK-12 learning.

Existing Tool/Software/Technology

The ESTEEM teacher education curriculum materials (project website: https://www.fi.ncsu.edu/projects/esteem/) are an example of a new resource that can be used to prepare K-12 teachers to teach data science. The materials emphasize preparation to teach data investigations that utilize the software program CODAP (https://codap.concord.org/) to improve students' data fluency.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 1)

This is an example of the State Frameworks we developed for the state of Florida, which we are piloting with a school district in North Florida as part of a grant. We have developed a four course program of study for the Career and Technical Division for grades 9-12 and have it aligned to a a draft version of the pending frameworks for the State/Community college Frameworks for Data Science and Fintech. We have started to develop an open source curriculum portal in Canvas under a CC-Share Alike license for use by teachers. We provide professional development and coaching to teachers to help them develop the skills they need to successfully teach this course. We are creating Data Science projects with Florida Themes in agriculture, health care, science, marine health, manufacturing, finance and other career related areas working with industry partners for data sets. We can provide expanded versions of these frameworks.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 2)

Program Title: Data Science
Program Type: Career Preparatory
Career Cluster: Information Technology
Teacher Certification Refer to the Course/Program Structure section.
This program offers a sequence of courses that provides coherent and rigorous content aligned with challenging academic standards and relevant technical knowledge and skills needed to prepare for further education and careers in Data Analytics and Data Science-enabled careers; provides technical skill proficiency, and includes competency-based applied learning that contributes to the academic knowledge, higher-order reasoning and problem-solving skills, work attitudes, general employability skills, technical skills, and occupation-specific skills, and knowledge of all aspects of AI and Machine Learning required for such data professionals working in business and academic environments. The intention of this course is to prepare students to be successful both personally and professionally in an increasingly data-focused society that demands more thorough data understanding and fluent analytic skills.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 3)

The content includes fundamental understanding and application of data analytics, data visualization, relational database design, machine learning, societal impacts of AI/ML, AI/ML systems and their components, problems and tools AI-enabled workers use to build models and systems that leverage data to make decisions, and mastery of foundational skills required to become power ML/AI users. In addition, the course content includes but is not limited to practical experiences in AI/ML system design, deployment, and evaluation; problem identification; creation, selection, and curation of data sets; computer programming, use of machine learning algorithms, program design structure, evaluation of the societal impact of AI, employing ethical and responsible development methodologies and decision making, essential programming techniques, and implementation issues. Specialized programming skills involving advanced mathematical calculations and statistics are also integrated into the curriculum.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 4)

New Course: Foundations of Programming for AI

01.0 Explain and use design thinking to solve a problem
02.0 Develop an awareness of microcomputers and how they work
03.0 Know the basic structure of a CPU (Central Processing Unit) and GPU (Graphics Processing Unit)
04.0 Explain what an ALU (arithmetic logic unit) is and how it works.
05.0 Demonstrate understanding about various aspects of digital memory.
06.0 Develop awareness of computer languages, web-based & software applications, and emerging technologies related to

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 5)

AI and Data Science

07.0 Explore the characteristics, tasks, work attributes, options, and tools associated with a career in data science.
08.0 Demonstrate an understanding of the characteristics, use, and selection of numerical, non-numerical, and logical data types.
09.0 Distinguish between iterative and non-iterative program control structures.
10.0 Describe the processes, methods, and conventions for software development and maintenance for data science.
11.0 Explain the types, uses, and limitations of testing for ensuring quality control.
12.0 Create a program design document to support engineering design of a program or product.
13.0 Create programs that solve a problem using non-iterative and iterative algorithms.
14.0 Design a computer program to meet specific physical, operational, and interaction criteria for data science.
15.0 Create and document a computer program that uses a variety of internal and control structures for manipulating varied data types.
16.0 Differentiate among procedural, object-oriented, compiled, interpreted, and translated programming languages.
17.0 Create and document an interactive computer program that employs functions, subroutines, or methods to receive, validate, and process user input or data sets.
18.0 Be able to read and write data (file I/O) to and from a program.
19.0 Solve problems using critical thinking skills, engineering design, creativity and innovation.
20.0 Describe the importance of ethical and fair use, security and privacy information sharing, ownership, licensure and copyright of created programs and data.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 6)

New Course: Data Analytics and Database Design

21.0 Generate and tell stories with data.
22.0 Think critically about data.
23.0 Collect, analyze, and visualize a dataset to gain insight.
24.0 Know and apply best practices in data visualization.
25.0 Construct interactive dashboards following best practices.
26.0 Understand and apply concepts in probability.
27.0 Understand and apply concepts in basic statistics.
28.0 Understand and apply concepts in statistical sampling.
29.0 Understand and apply concepts in hypothesis testing.
30.0 Be aware of the limitations of statistics and cognitive biases in data analysis.
31.0 Understand how data is accessed, sorted, and stored.
32.0 Become more SQL literate.
33.0 Understand design considerations and apply best practices in designing SQL databases.
34.0 Identify data privacy/ data governance issues and methods to mitigate them.
35.0 Understand common cybersecurity issues and mitigation methods.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 7)

New Course:Foundations of Machine Learning and Applications:

36.0 Identify and define intelligent behavior.
37.0 Articulate the relationship between AI, Machine Learning, and Computer Science.
38.0 Identify and describe the types of representations and algorithms designed into AI-enabled technologies.
39.0 Define and investigate examples of AI applications.
40.0 Explain how domain knowledge is used in the design of AI systems.
41.0 Describe machine learning algorithms in AI-enabled technologies.
42.0 Explain the key technical challenges in design and responsible use of AI technologies.
43.0 Describe different types of data and how they are used in AI.
44.0 Explain and use design thinking to solve a problem.
45.0 Apply the machine learning life cycle in the development and use of a machine learning model.
46.0 Recognize and identify mathematical principles upon which machine learning and AI are built such as calculus, linear algebra, probability, statistics, and optimization partial derivatives.
47.0 Train and evaluate a range of ML models based on specific accuracy, inclusivity, and ethical design criteria.
48.0 Use and evaluate supervised learning techniques to classify or predict outputs.
49.0 Use and evaluate unsupervised learning techniques to solve problems.
50.0 Understand neural networks and their components.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 8)

New Course:Foundations of Machine Learning and Applications:

51.0 Use and evaluate different types of neural network architectures and their applications.
52.0 Use and evaluate reinforcement learning techniques to solve problems.
53.0 Research and explain the advancements in computing hardware that make AI possible.
54.0 Understand and articulate how AI can impact society in both positive and negative ways.
55.0 Understand the best practices and key characteristics of bias, fairness, transparency, explainability, accountability of ethically designed AI systems and decision-making practices.
56.0 Identify different kinds of data, their sources, and how they might be used in decision making.
57.0 Critique data and data-based claims to avoid being misled by data through identifying bias, confounding, and random error.
58.0 Describe the limitations of machine learning and the decisions that can be made with data.
59.0 Explore the characteristics, tasks, work attributes, options, and tools associated with AI-enabled careers.
60.0 Identify how leadership development, school and community service projects and competitive events are integral parts of career and technology education.

Existing Tool/Software/Technology

State Frameworks for Florida State (Part 9)

Capstone Project:

61.0 Explain and use design thinking to solve a problem.

62.0 Design AI solutions using embedded computing (as applicable to specific projects).

63.0 Explore the characteristics, tasks, work attributes, options, and tools associated with AI-enabled careers and educational pathways to achieve these career goals.

64.0 Use appropriate tools to design an AI System to solve problems.

65.0 Set up and use a machine learning (ML) pipeline to solve a problem.

66.0 Appropriately use automated services to accomplish common tasks.

67.0 Use data analysis and visualization tools to work with datasets and gain insights.

68.0 Apply the machine learning life cycle in the development and use of a machine learning model.

69.0 Design and develop AI systems to solve a problem or design solutions for social and ethical issues.

70.0 Create a portfolio of AI projects that demonstrate ability to program machine learning models using a wide range of AI algorithms.

71.0 Research and evaluate various AI careers involved in AI system usage, design, development, deployment, and maintenance.

Real World Application

Strengthening Data Literacy across the Curriculum (SDLC) (Part 1)

To promote understanding of and interest in data literacy among students from historically marginalized groups, the Strengthening Data Literacy across the Curriculum (SDLC) project developed and has been studying high school mathematics curriculum modules that focus on social justice issues (https://sites.google.com/view/uss-data/home). One module supports student analyses of income inequality in the U.S. using U.S. Census Bureau microdata and the online data analysis tool CODAP. In fall 2019, the module was implemented by seven teachers of 12th-grade non-Advanced Placement mathematics classes in six Northeast schools with high proportions of students from Black, Latinx, and low-income families. Based on pre- and post-module assessments and results from almost 200 students, we found statistically significant growth in students' understanding of important statistical concepts and interests in data analysis. We also found signs of greater social and political awareness and agency with data – outcomes associated with increased critical data literacy.

Real World Application

Strengthening Data Literacy across the Curriculum (SDLC) (Part 2)

In this module, students work through seven lessons and a final team data investigation, exploring different forms of income inequality, its scope, and possible explanations. Students investigate what the wages of high-, middle-, and low-income earners in the U.S. have looked like over time, using random samples of 1,000 or more individuals drawn through a U.S. census microdata portal that was developed for CODAP. They also examine the male-female wage gap and whether it can be explained by a third variable. As students explore these issues, they deepen their abilities to compare quantitative distributions using measures of center and variability. They also build abilities to reason with multivariable data – a skill that should be developed in pre-K-12 education but is not currently emphasized (e.g., Engel, 2016).

We found that students' understandings of assessed statistical concepts grew between the start and end of the module at $p<0.0001$ and with a moderate effect size of $d=0.43$. In addition, students' interests in statistics and data analysis, measured through a pre- and post-module interest scale, grew significantly at $p=0.001$ with a small effect size of $d=0.25$. Qualitative data suggest that the module helped multiple students develop a better understanding of social conditions as well as agency to work toward social change. In an interview, one student shared: "I think we need to do more of [these modules] in school, to be honest with you. A lot of people are afraid to talk about issues like that… If more kids tackle it head-on, they'll have a better understanding of how our world really works." Another student shared: "It was just very interesting being able to graph out everything and finding all these differences and like, oh, this actually is an issue… People aren't just making things up. This is a real problem, and hopefully we can figure something out."

Existing Tool/Software/Technology

WeatherX (Part 1)

WeatherX is an NSF-funded project that has been developing curriculum strategies to promote scientific data practices and interest in data science careers among middle-school students in low-income rural areas. The project has iteratively developed and tested two multi-week curriculum units in which students develop fundamental scientific data practices as they learn about local weather phenomena and the climate.

A major tool that WeatherX has developed is the NOAA Weather portal, a permanent plug-in within the online data analysis and visualization tool CODAP. The portal provides anyone with a web browser access to large-scale weather data from the National Oceanic and Atmospheric Administration. Using the portal, individuals can select and download hourly, daily, and monthly weather measurements (such as for temperature, wind speed, and precipitation levels) from 1,783 weather stations in the U.S., with records spanning as far back as the mid-1800's to the present. Students using WeatherX lessons investigate data from their own local weather stations and make claims about whether specific weather events are extreme by comparing event data with 30-year climate averages. Because students in our project live in northern New Hampshire and Maine, they have also used data from the NOAA Weather portal to examine weather events on the summit of Mount Washington, New Hampshire – a site that has been called the "Home of the World's Worst Weather." Data help them see that extreme winds in their location may not be extreme on the summit of Mount Washington!

Existing Tool/Software/Technology

WeatherX (Part 2)

Based on research during an initial implementation of the WeatherX units, students and teachers spoke positively of the ability to investigate and learn from large-scale weather data from their own and other locations using the NOAA Weather portal and CODAP. In addition, teachers observed, and students reported, higher levels of confidence in making and interpreting graphs in CODAP after engaging in WeatherX activities. We will be making our WeatherX units available for public use in the upcoming months. In the meantime, anyone can explore vast troves of weather data – and how extreme or typical these data may be – from all over the U.S. using the NOAA Weather portal, available at codap.concord.org under the “plug-ins” menu.

An online tutorial on how to make graphs in CODAP to examine weather data is available at https://codap.concord.org/app/static/dg/en/cert/index.html#shared=https://cfm-shared.concord.org/vljl6HALM3Kwxorhb0EV/file.json.

Challenge Educators Face

One goal (an important one, in my opinion) in data science education is for students to pose investigative questions that are important to them, and to use data to answer (at least partially) their questions. This poses big challenges to the instructor, who must potentially grade or assess many different projects, and must have sufficient confidence in their own knowledge of data analysis to assist students to follow many different pathways, without knowing ahead of time what the "right" answer is. Even when working with the same data set, different students might pursue different approaches. How do we prepare instructors so that they can support students to follow valid approaches without teaching the misconception that "you can say anything with data"?

Real World Application

Using data cards for data exploration and unplugged introduction of decision trees with data with nutrition facts of food products. Children (grade 5 and 6) are asked to develop a multistep decision rule for predicting whether a food is recommendable or not. Children and sort the data cards, identify reasonable split criteria and experience the different rates of false classifications. The dependence of the decision rule on the somewhat subjective labelling and the training data set can be experienced.

Podworny, S., Fleischer, Y., Hüsing, S., Biehler, R., Frischemeier, D., Höper, L., & Schulte, C. (2021). Using data cards for teaching data based decision trees in middle school. In 21st Koli Calling International Conference on Computing Education Research (Koli Calling '21), November 18-21, 2021, Joensuu, Finland. ACM. https://doi.org/10.1145/3488042.3489966

Existing Tool/Software/Technology

Educationally designed Jupyter Notebooks based on Python make complex machine learning algorithms accessible and transparent to students. Jupiter Notebooks can be designed with a kind of menu-driven interface, where coding and recoding parts can be used but that is not necessary. Jupyter Notebooks can be designed as worked examples that can scaffold students in writing computational essay for their own data analysis. A computational essay can be conceived as in interactive changeable book with, code, graphs, tables and text that makes a data analysis reproducible.

Fleischer, Y., Biehler, R., & Schulte, C. (2022). Teaching and Learning Data-Driven Machine Learning with Educationally Designed Jupyter Notebooks. Statistics Education Research Journal, 21(2). https://doi.org/10.52041/serj.v21i2.61

Real World Application

This paper reports on progress in the development of a teaching module on machine learning with decision trees for secondary-school students, in which students use survey data about media use of youth to predict who plays online games frequently (to suggest specific advertisements for online games to these students. The data set consists of more than 500 cases and more than 50 variables and is also used for just data exploration activities as a preparatory step before introducing machine learning.

This context is familiar to students and provides a link between school and everyday experience. In this module, they use CODAP's “Arbor” plug-in to manually build decision trees and understand how to systematically build trees based on data. Further on, the students use a menu-based environment in a Jupyter Notebook to apply an algorithm that automatically generates decision trees and to evaluate and optimize the performance of these. Students acquire technical and conceptual skills but also reflect on personal and social aspects of the uses of algorithms from machine learning.

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. Teaching Statistics, 43(S1), S133-S142. https://doi.org/10.1111/test.12279

Real World Application

In this paper, we will describe an introduction to Data Science for secondary school students. We will report on the design and implementation of an introductory unit on “Data and data detectives with CODAP” in which secondary school students used the online tool CODAP to explore real and meaningful survey data on leisure time activities and media use (so-called JIM-PB data) in a statistical project setting as a starting point for data science. The JIM-PB data set served as a valuable data set that offered meaningful and exciting opportunities for data exploration for secondary school students, and CODAP proved to be a valuable tool for the first explorations of this data.

The JIM-PB data set has more than 500 cases and more than 50 variables. The related questionnaire is similar to one that is used for representative surveys of youth's media use, which is published biannually. The published report also contains aggregate data, whereas our data set has raw micro-data and students can relate their own exploration to the results of the national report. The students can use the questionnaire to collect their own class or school data. The results can be used to reflect the classical and social media use of the students. Based on their everyday knowledge, they can generate hypotheses and interpretations of results.

Frischemeier, D., Biehler, R., Podworny, S., & Budde, L. (2021). A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. Teaching Statistics, 43(S1), S182-S189. https://doi.org/https://doi.org/10.1111/test.12283

Real World Application

NetApp Data Explorers (Part 1)

NetApp Data Explorers is an after-school program, funded by NetApp, that introduces middle-school students to data science in the context of “using data for social good.” Data Explorers focuses on the 17 United Nations Sustainable Development Goals (SDGs), which present a “blueprint to achieve a better and more sustainable future for all” and introduces participants to the data the UN has collected to track progress in the SDGs. Using CODAP (Common Online Data Analysis Platform, an educational data visualization tool), students explore health and education indicators from 195 countries that belong to the United Nations. The attributes include life expectancy (overall, for males, and for females), average years of school attended (overall, for males, and for females), teen birthrate, population in millions, percent of the population that is urban, medical doctors per 100,000 people, and percent of the population who use the Internet.

Real World Application

NetApp Data Explorers (Part 2)

Students then analyze more local data (county-level data in the US and similar administrative structures in other countries), focusing on health and education. In the US, the attributes include a set of health outcomes (e.g. life expectancy), health factors (e.g., % of people who smoke and medical doctors per 100,000 people), social and economic factors (e.g. median income and high school graduation rate) and racial demographic data. In the final step in the Data Explorers program, teams of students carry out a project "digging deeper" into an attribute they find particularly interesting, then creating a presentation describing their data-based discoveries and issuing a call to action, either to find out more about the situation, take action to improve it, or increase awareness around it.

While Data Explorers has been very engaging for participants in its pilot phases in the US, England and the Netherlands, we have also identified some challenges in the process of implementing (and revising) the program. For example, while middle-school students are especially attuned to issues of "social good" and inequity, it is a big leap from noticing, for example, that African countries have lower life expectancies to coming up with a sensible and realistic "call to action." In addition, much of the data is reported as rates (e.g. number of doctors per 100,000 people in a country), which may be mathematically challenging for some students to understand. Finally, all of these kinds of "civic data" are aggregated across a country, county, or other political entity; this kind of aggregation disguises a lot of local variability, so runs contrary to many youths' experience, which tends to be at much smaller geographical scale, like a neighborhood.

Challenge Educators Face

EDC's Oceans of Data Institute (Part 1)

In our work at EDC's Oceans of Data Institute, one of our areas of interest has been in exploring how educators can integrate the use authentic data into their teaching practice. Topic areas in which we've developed and tested curricula include Earth science, biology, social studies, and civil engineering. In all cases, we've had student work with what we call "CLIP" data:

- Complex – having multiple variables within the dataset;
- Large – with more data than are required to address any single question or problem;
- Interactively accessed – meaning that students visualize and analyze data using a computer;
- Professionally-collected – often by agencies or scientists, but always including data from someone other than the individual student.

This approach models the real-world situations students are likely to face when using data, and helps to build those critical "habits of mind" of identifying the question they are trying to answer using data, determining which data within the dataset are relevant to their question, and then carrying out whatever analyses are required to gain the relevant information they are seeking.

Challenge Educators Face

EDC's Oceans of Data Institute (Part 2)

One of the recurring challenges in having students work with CLIP data has to do with the nature of the interface itself, which typically has to do three things simultaneously. It must:

- Provide the opportunity for students to explore and manipulate the data in various forms, often including tables, graphs, and maps;
- Allow space for written instructions, prompts, or background;
- Give students a way to capture their work – ideally in such a way that the instructor can review it and provide feedback.

Over the past several years we have tried a number of different solutions to this problem. In the NSF-funded ZoomIn! project, we used a bespoke Web-based interface for instruction and student work, along with an expandable data interactive window housing Concord Consortium's excellent CODAP interface. In the NASA-funded Real World, Real Science project, we used Concord's LARA framework, which also provides space for instruction and student work, and which easily integrates either CODAP or SageModeler interactive data windows.

Challenge Educators Face

EDC's Oceans of Data Institute (Part 3)

More recently we've also integrated CODAP and SageModeler data windows into Canvas – a commercial learning management system. Although Canvas would allow the integration of these tools within a Canvas page, the fact that the instructional window is often surrounded by other website navigation, etc. (thus reducing available screen real estate), we've more often had students launch the data interactive window in a separate browser tab and had them switch back and forth between tabs to complete their work. We are currently testing this approach with undergraduates and early feedback suggests that most students are comfortable managing multiple windows simultaneously; but we are curious to know whether this is a viable approach for younger students, and if so, what the youngest age is where it would be viable.

Real World Application

Data Jam/Puerto Rico

Data Jam/Puerto Rico is a 5-year project in Puerto Rico to introduce middle- and high-school students to data science by engaging them in authentic environmental investigations using long-term ecological datasets. Students engaged in Data Jam/Puerto Rico are provided access to decades of long-term ecological data from El Yunque National Forest gathered by the Luquillo Long-Term Ecological Research (LTER) program at the University of Puerto Rico. After a series of introductory activities, students work in teams to come up with a research question, analyze the data using the Common Online Data Analysis Platform (CODAP) and use the evidence from their analyses to answer their research question. Some of the groups present their posters at an annual Data Jam Symposium.

The project team's primary activities are providing professional development for the participating teachers (most of whom teach either science or math), supporting teachers in classroom implementation through a combination of educational materials and classroom visits, and curating data sets from the LTER so that they are possible for students to use. In this last task, the team has been supported by a data fellow funded by the Environmental Data Initiative (EDI).

The intervention also includes several scientific mentors, young Puerto Rican scientists who serve as role models for students by sharing their career paths and advising them on their data analysis.

Challenge Educators Face

EMBEDS (Exploring the Mathematics of Biological Ecosystems with Data Science) (Part 1)

The EMBEDS (Exploring the Mathematics of Biological Ecosystems with Data Science) project at TERC and the University of Colorado, Boulder is integrating “data excursions” into a ninth grade biology unit that focuses on the changing ecology of the Serengeti. The intent is to have students use the same datasets scientists used to understand why the populations of both wildebeest and buffalo increased rapidly and substantially between 1960 and 1975. Several challenges arose. First, scientific datasets from several decades ago are not always available; if they are, they likely require a lot of work to make them usable by current-day students. Second, techniques for measuring ecological quantities are notoriously difficult – and were much less precise in 1960 than they are now. In our case, students wanted to determine if increased food could have been the driving force behind the increase in animal populations. How do you measure the amount of available food in an area like the Serengeti?? After some research, we discovered that scientists didn’t know how to measure that quantity either – and they used rainfall as a proxy. So our data search turned to rainfall data for that period, which we did find and provided to students. While these data indicated that rainfall had not increased during the period of population growth, drawing the conclusion that a change in food availability did not cause the population explosion required students to do two difficult things: 1) understand how data can indicate NO relationship between two variables and 2) reason about food availability through the proxy quantity of rainfall.

Challenge Educators Face

EMBEDS (Exploring the Mathematics of Biological Ecosystems with Data Science) (Part 2)

These challenges arose because we were trying to fulfill two different learning goals at once: 1) have students learn some basics of data science and 2) have students learn particular science concepts dictated by the NGSS. While the NGSS suggests that students in ninth grade use mathematical and computational models in the pursuit of science, they are vague about what the process entails – and don't acknowledge that simultaneously fulfilling the dual goals of science and data fluency may not be straightforward. In our case, we found that students hadn't had much experience looking at covariation – which was necessary to explore the relationship between rainfall and population – so we needed to develop materials to introduce basic graph analysis, as well as to the relevant scientific principles.

Real World Application

FINANCIAL LITERACY - RETIREMENT CALCULATIONS

Students decide on a retirement age (my students used age 65) to calculate their age until retirement. They decide on 2 amounts they think they could save per week - one realistic, one a stretch and everyone uses those amounts for comparison purposes. They use TheMint.org's compounding calculator to figure amounts until retirement if that weekly money (calculated to yearly savings) is invested in a savings account (.50% return) or the stock market (7% return - historically). Next, they collect and share that data in either an online spreadsheet or a paper graph. Students make predictions about how weekly savings amounts translate across time and investment vehicle. Some students multiply by 48 weeks instead of 52, so there are opportunities for conversations there. (Also requires conversations around unpredictability of stock market as well as how fees, buying and selling, etc. affect growth.) Students who initially said they had NO money to invest, usually are thinking by the end of the activity about how much they could spare per week. MATH STANDARD: Represent and analyze quantitative relationships between dependent and independent variables. MATERIALS: Kids Compounding Calculator at TheMint.org; Excel, Sheets or a paper version of a chart; Chromebooks; paper and pencil. EXTENSION ACTIVITY: Calculate how much money a person could save if they had a better credit rating on a vehicle loan or a house mortgage. (Use an online loan calculator.) Take the difference in those two amounts and have students see how much that could equate to if it was invested and compounded over time. (In our class example, the DIFFERENCE paid out over the life of a 30-year mortgage ($160,000) - invested instead - equaled over $500,000 that the person would have been able to save by the time the mortgage was paid off.)

Challenge Educators Face

Teachers (at all educational levels, K-college) often struggle to incorporate real-world, complex, "messy" datasets into classroom instructional activities - even when those data are easily available in online sources, like interactive data maps or news features. Students easily navigate to confusing and complicated places in those data sources, and a teacher can be utterly lost when coming around to help them.

Existing Tool/Software/Technology

The website SocialExplorer (https://www.socialexplorer.com/), developed by sociologist Andrew Beveridge, is an excellent US census data access tool. It is useful for researchers, teachers, students, and the public. The subscription version provides access to historical census data back 1790, and the free public version has great mapping and reporting tools for recent censuses.

Real World Application

DATAJAM https://www.pghdataworks.org/data-jam (Part 1)

Pittsburgh DataWorks, along with the West Big Data Innovation Hub and the New York Hall of Science developed a virtualized version of the DataJam, an annual program and competition for high school students that runs throughout the academic year and introduces youth to data visualization and analytic skills to answer a research question of their own design. Students are guided by DataJam mentors, who are volunteer university undergraduate students trained in mentorship skills as well as statistics, data ethics and community-based research principals. A large number of resources for DataJam teams have been developed and are available for free on the DataWorks website. Teams present their findings both in a poster and an oral presentation; judges are from academia and industry. Field trips introduce youth to a wide variety of careers in which data science skills are valuable.

Real World Application

DATAJAM https://www.pghdataworks.org/data-jam (Part 2)

The COVID-19 pandemic forced us to migrate DataJam to 100% online for the 2021-22 school year. We used this opportunity to pilot the feasibility of using the online format to expand the program to culturally and geographically diverse teams and an eMentoring model. This provides us with an opportunity to determine how to expand, scale and replicate the success of DataJam to reach the most underserved communities that lack access to the resources and expertise to participate in rich data science programs like DataJam in person. In the pilot project we expanded DataJam participation to high schools across Pennsylvania, as well as New Jersey and Massachusetts through an eMentoring approach using videoconferencing and other online collaboration tools to support high school research teams and teachers. In collaboration with the West Big Data Innovation Hub a student team from the Pala Native American Tribal Lands in Southern California also participated. Focus groups and surveys conducted with the students and teachers indicated that the program effectively translated to function online in very diverse cultural and geographical settings. That modest success leads us to explore expanding the diversity of high school participants to include underserved urban, rural, immigrant, unhoused, and tribal youth, and to deepen our understanding of the factors that lead to changes in self-efficacy and career aspirations in the target youth in this online version of the program.

Real World Application

My background is Pediatrics, Immunology and Medical Informatics. I have a profound experience of using ontology for data integration and data analysis in health and regulatory area. I have been mentoring junior high schoolers and college students to work on FDA funded informatics projects or University collaborative research projects. Out of 4 female (all minority background) students, 3 switched their major or chose the major in Data Science or IT. One of them switched from pre-med to data science track. I keep mentoring those students are currently in college majoring data science for continuous data science research.

Challenge Educators Face

The NSF-funded InquirySpace projects, aimed at providing opportunities for helping youth engage in independent STEM investigations, represent a successful real-world application of data science. Aspects of the curriculum provide enlightening opportunities that other programs may be able to adopt or emulate. One specific set of examples derives from classroom observations in the project that showed students leveraging graphs as epistemic tools. Video analysis of extended classroom interactions indicated that student use of data graphs during sensor-based experiments could range from a focus on producing them as a procedural display to engaging deeply over several days to make sense of different graphical representations of their data. In a multi-case study, we investigated high school physics students' use of graphs and what prompted those uses. Unexpected data patterns and graphical anomalies appeared to play a strong role in provoking student reasoning and triggering engagement with the graphs. For the three representative groups subject to in-depth analysis, graphs appeared to play an important role in their knowledge production as they made decisions about their experimental procedures and goodness of data. We conclude that when students produce data in an experiment for which they feel a sense of ownership, they can exhibit an almost proprietary interest in representations of their data. As they work to create alignments between their conceptions of their data and the unexpected data patterns they see in the graphs, they begin to use their graphs as epistemic tools.

Real World Application

The NSF-funded StoryQ project, aimed at providing opportunities for youth to use data science and artificial intelligence (AI) to explore language-centered scenarios, represent a successful real-world application of data science. One example of this application engages students in using a data-focused view to analyze poetry. Students process poetry to identify and characterize features of its language such as the number and type of syllables in each line, then use patterns uncovered in the data to develop and train a machine learning model in the StoryQ app, a plugin to the Common Online Data Analysis Platform (CODAP). The goal of the curriculum is to scaffold text analytics for ELA students so they learn to understand and appreciate both poetry and applications of data and AI. Data analysis lies doubly at the heart of the StoryQ activities, as students not only use analysis to examine and identify patterns in language to train a model but also use data visualization and analysis to evaluate the results of the model, directly evaluating the weights of various parameters derived from applying their model in order to characterize the text itself as well as to judge the accuracy of the model. In this manner, StoryQ stands as an example of the ways students can use data analysis as a tool both to shed new light on traditional school subjects and as a pathway toward understanding new applications of data and computing.

Existing Tool/Software/Technology

DataFlow (Part 1)

The DataFlow environment, originally developed via an NSF-funded project aimed at integrating computational tools and computational thinking practices into the high school life science curriculum, stands as a software tool that could broadly supplement data science learning. Developed as part of InSPECT, a project aimed at engaging students directly with producing and processing data directly.

DataFlow is a comprehensive open source platform for programming, data processing, and real-time data graphing. Within DataFlow, students are able to produce meaningful data and control their data through its lifecycle, making decisions as data flows from the collection device to a representation on screen. Students choose what data is being collected, where to collect it, how to modify or transform it, how to use the data to actuate a relay, how to store that data, and how to view the data.

To program in DataFlow, students drag and connect nodes in an open central workspace. For example a student may use a CO2 sensor node and a number node as inputs to a logic node in order to to construct an if-then statement that controls a relay. The student can also connect the CO2 sensor and light sensor nodes to a data storage node, which allows them to view a real-time graph of their data when the program is run. Visualized wires connecting DataFlow nodes make the the data pathway visible: data flows from one node to another until it reaches an end point.

Existing Tool/Software/Technology

DataFlow (Part 2)

Unlike traditional block programming, students choose how to orient nodes in the Dataflow workspace. For example, if a student wants to read their program from right to left, they could position their nodes in that fashion. A node selector in the left-hand column of the workspace offers students additional programming options. A timer node actuates a relay based on a time schedule determined by the student. The generator node generates sine, square, and triangle waves with controllable amplitudes and periods. The math and transform nodes perform operations on the desired input. This wire and node programming style gives students the opportunity to view their program as a pipeline.

DataFlow exposes students to essential concepts of data production and processing, enabling them to engage with hardware and software that work together in an Internet of Things (IoT) system. This opens many new pedagogical possibilities, as students can monitor CO2 levels in the classroom next door or one thousands of miles away. By enabling understanding of IoT systems, DataFlow demonstrates how data can move beyond the confines of the classroom and how data can open opportunities to control and interact with the world directly.

Existing Tool/Software/Technology

When working with real-world data, teachers and curriculum developers face many dilemmas. One of the most prominent lies in the complexity of these datasets. For example some relevant datasets in ongoing projects may include tens of thousands of cases and over one hundred attributes. These datasets on their own are much too large and complicated for the middle school audience the project targets. Choosy, a simple data tool designed to address this issue, could be adopted broadly to enhance data science curricula and approaches.

Designed to enable the straightforward creation of simple datasets from complex ones, Choosy is implemented as a plugin to the Common Online Data Analysis Platform (CODAP). By allowing teachers or curriculum developers to work with and easily select sub-portions of larger datasets, it enables classroom activities or curriculum units to involve a much wider range of datasets than otherwise possible. Choosy has a simple interface, with a tab for selecting the dataset, a tab for attributes, and a tab for tagging specific cases. Users proceed through these tabs, using a dropdown to select their desired dataset, a series of checkboxes to identify the attributes they desire to include, and a streamlined interface to select specific cases to set aside or delete. When they are satisfied with their choices (which can be monitored in a live readout reporting the number of attributes and cases currently selected), users can export the final simplified dataset for inclusion in their desired activity.

Existing Tool/Software/Technology

In using the Common Online Data Analysis Platform (CODAP), student can easily make simple changes to datasets by dragging and dropping attributes and creating formulas. However, this streamlined interface can render more complex operations difficult or impossible at times. Additionally, changes to datasets must be performed individually and repeated by hand, making the application of a series of data moves on multiple datasets tedious or prone to error.

The Transformers plugin for CODAP is a tool aimed at addressing this issue. Designed by the Bootstrap curriculum with the goal of featuring data moves more prominently in their curriculum resources that use CODAP, the plugin provides 30 different transformations for users to apply to datasets automatically. Among many other transformations, this set of built-in transformations makes it easier for students to filter and sort attributes and to measure, aggregate, and summarize data. Additionally, the tool enables a more programmatic, documented and repeatable approach to data transformation within the CODAP environment. Users can save sets of transformations as miniature programs, coming back to them later or saving them as mini-tools for future use.

Additionally, the tool provides specifically for simple experimentation and iterative learning. Users can transform datasets to produce new, distinct output datasets without modifying the original input dataset itself, thus enabling easy “what if” exploration and comparison of datasets that may represent distinct transformations performed on the same source dataset.

Existing Tool/Software/Technology

Students facing complex datasets can often struggle to make sense of them. Patterns buried deep within the datasets can be difficult to tease out, and the existence of multiple parameters can make discerning nuanced patterns complex. All of these problems are common for students examining data visualizations. However, a significant fraction of learners are blind or visually impaired, rendering traditional visual data analysis tools practically inaccessible. An experimental plugin to the Common Online Data Analysis Platform (CODAP) aims to begin to address both of these issues by making data “visible” through signification, the use of audio portrayals of data.

The signification plugin for CODAP represents an experimental approach to allowing users to explore data through audio representations. It includes an interface for selecting a sub-portion of a dataset and a play button that loops through the selected subset repeatedly. As it loops through the dataset, the plugin translates the data into audio signals, varying aspects of the audio including pitch, timbre, and other attributes. Users can assign different attributes of a dataset to different aspects of the audio output to allow for examination of multiple attributes simultaneously, modify the speed and repetition rate of the looping to enable closer study of patterns, and easily compare portions of a dataset with each other to make differences more readily apparent.

While this work is still in the prototype stage, early experiments with both visually impaired and fully sighted learners make its power very clear. Sonifying data can make subtle patterns more detectable and can otherwise provide a “new lens” on data that has already been examined visually. Many scientists working with complex datasets in the laboratory use signification as an additional tool in their data analysis toolbox — it’s time for data science education to begin to explore its possibilities as well.

Existing Tool/Software/Technology

Part 1

Examining complex datasets is like a journey. New ideas arise in the process of exploration, leading to blind alleys or opening up new twists and turns. One may find something interesting that is important but separate from the initial question, creating the desire to set it aside for later examination. Or one may encounter several different aspects of the dataset that are not striking individually but stand as clearly important when considered in concert.

This journey is a common aspect of data exploration, but poses a series of hairy dilemmas. When learners wish to explore an unexpected finding at a later point, they may be unable to bring the data back to that state easily. Reviewing data from multiple angles can be inherently challenging. Most importantly, many tools limit students to producing a visualization as an endpoint, reducing the nuanced journey of data exploration into a small group of finalized visualizations. The Story Builder plugin for the Common Online Data Analysis Platform (CODAP) is designed to provide learners with the ability to convey rich “data storytelling” experiences easily.

Existing Tool/Software/Technology

Part 2

Inspired by the NSF-funded Writing Data Stories project's goal of helping middle school students become "data storytellers," the Story Builder plugin grew out of the desire to allow learners to use data as a medium to express their understanding of important socio-scientific issues. The Story Builder plugin, designed by Tim Erickson and programmed further by Bill Finzer, allows students to build story "moments," with each interactive moment capturing the state of a CODAP document at a given time. Since CODAP can also embed web pages and videos, a story can be truly multimedia. The plugin allows students to edit, delete, and rearrange their "moments" or lock them to prevent accidental changes.

Initial work to date has shown that stories created with the Story Builder plugin are great for student projects and presentations, introductions to data-rich content, arrangements of data sequences for curriculum development and even applications as broad as a prototype mockup environment for designing new CODAP plugin capabilities. The Story Builder plugin stands as a useful example of an affordance that can add important new dimensions to data science education applications, and we encourage others to explore and help build on its capabilities.

Existing Tool/Software/Technology

Data Story Bytes or "DataBytes" describes a framework for discussing data visualizations that may appear as part of a teacher's curriculum, or in local or national news stories, with their students. These activities are designed to support quick (30 mins) discussions, in the spirit of “number talks,” to critically analyze and interpret data visualizations in ways that connect to students’ lives and important issues in society. Teachers can select from a series of questions designed to guide students through:

- Making sense of trends and relationships in the data or visualization, what these patterns mean, and how they connect to key science concepts.
- Building personal connections by considering how students’ own lives and communities may be impacted by or reflected by the patterns found in data.
- Reflecting on the context and history of the data, how it was collected, by whom (including what gets “counted” and why), how it is visualized, what might be missing/hidden, and what questions the data can and cannot answer.
- Envisioning future uses of data and visualization to expand the investigation, include and explore different perspectives, and highlight the importance of understanding what’s happening in the world around us in multiple ways.

Our student-facing materials (formatted as slide decks, see here or see the links in the teacher guides below) are fully bilingual (English/Spanish). This document also comes with a toolkit for teachers to build their own DataBytes activities, a student glossary of key terms in English and Spanish, and teaching ideas for supporting multilingual students.

Real World Application

Former and current biology students at Maize High School conducted primary water quality sampling in partnership with the United States Geological Survey - Kansas (USGS-KS) in 2018 to develop a predictive model for the occurrence of harmful algal blooms (HABs) in Cheney Reservoir, the primary drinking water supply for Wichita, Kansas and surrounding communities. Using HOBO temperature and light sensors on a thermistor chain to collect data in the reservoir, they analyzed their primary data using rLakeAnalyzer and Concord Consortium's Common Online Data Analysis Platform (CODAP) in tandem with Kansas State University's Kansas Mesonet wind data. Students and their teacher found that Cheney Reservoir, which was previously believed to not stratify, does indeed have low/no wind periods that lead to short-lived thermal micro-stratifications. When wind speeds again increase, polymictic mixing begins again, allowing for nutrients that were once at the bottom of the reservoir and unobtainable for use by surface algal colonies and cyanobacteria to be moved up to the surface of the reservoir, within the photic secchi depth, favoring the formation of HABs. The resulting HABs not only costs drinking water treatment systems but also risks exposing the public to potent neurotoxins and hepatotoxins. However, predicting their occurrence by monitoring both wind patterns and hypoxia / anoxia at the bottom of the reservoir curbs these costs for all stakeholders.

Real World Application

As a result, on October 13, 2020, President Trump issued an Executive Order (EO) 13956, titled “Modernizing America's Water Resource Management and Water Infrastructure”, bipartisan Congressional legislation was passed to develop this new data-driven approach to proactive water treatment, and USDA-NRS and several commercial agriculture entities adopted land around Cheney Reservoir to begin a multiyear, regenerative agriculture, soil health initiative -- all being scaled across the U.S. at this time. The students ALL became STEM majors at universities across the U.S., and the teacher was contracted by Microsoft as a big data analyst. She wrote FarmBeats for Students machine learning and artificial intelligence curricula for Future Farmers of America (FFA) and Center for Agriculture Science Education (CASE) and helped to develop the machine learning methodologies being deployed across all industries. Subsequently, NOAA, under the umbrella of U.S. Department of Commerce, funded the Alabama Water Institute (AWI), and the teacher is now the Director of Regional and National Collaborations at AWI, developing collaborative work for NOAA’s CI for Research to Operations in Hydrology (CIROH) and USGS-Hydrologic Instrumentation Facility in partnership with the National Water Center in Tuscaloosa, AL. Working in tandem with the Department of Defense-supported Global Water Security Center (GWSC), the teacher and her growing body of students aim to continue to develop the water intelligence for this new water economy.

Real World Application

In our state, we had a success in Data Science with youth as we developed a course for high schoolers based upon data science, and offering a course that students could receive either math or computer science credit. A lot of time and research went in to developing this course, and we believe it highlights the changes in education, and the importance of CS and Data Science in today's education.

Existing Tool/Software/Technology

The government is making large strides toward making open data sets findable and accessible. Along with this is a growing number of solutions to make those data sets usable. Each solution targets a different audience and, therefore, differs with regard to underlying technology, capabilities, data access strategy and requisite skills for use. In the past, I have proposed a game to teach critical thinking approaches to problem solving. The game targets K-12 students and is based on the scenario discussed in this example . Grafana (https://grafana.com/) is open-source software that is suitable for both the needs of scientists and K-12 students as envisioned above. The primary function of Grafana is logging, but it can be adapted to build the game described above.

Real World Application

The Next Grand Challenge: Building Critical Mass for Data Science

The *California Education Learning Lab* is pleased to announce the release of a new grant opportunity to promote the buildout of critical data science educational infrastructure. Through this RFP, Learning Lab's Grand Challenge seeks to incentivize public higher education institutions to embrace data science as an opportunity to build new pathways, modernize majors, attract historically underrepresented students into STEM, and deepen both civic and interdisciplinary learning.

Interested in Applying?

Review the guidelines for this grant opportunity and **submit your Statement of Intent** using our Online Application Portal by **Friday, December 9, 2022**.

https://calearninglab.org/grant/data-science-rfp/