The Mathematics Enthusiast

Manuscript 1573

# A Mathematics Pipeline to Student Success in Data Analytics through Course-Based Undergraduate Research

Kristin P. Bennett

John S. Erickson

Amy Svirsky

Josephine C. Seddon

# A Mathematics Pipeline to Student Success in Data Analytics through Course-Based Undergraduate Research

Kristin P. Bennett, John S. Erickson, Amy Svirsky
Rensselaer Polytechnic Institute, United States

Josephine C. Seddon
University of Rochester, United States

### Abstract

This paper reports on *Data Analytics Research* (DAR), a course-based undergraduate research experience (CURE) in which undergraduate students conduct data analysis research on open real-world problems for industry, university, and community clients. We describe how DAR, offered by the Mathematical Sciences Department at Rensselaer Polytechnic Institute (RPI), is an essential part of an early low-barrier pipeline into data analytics studies and careers for diverse students. Students first take a foundational course, typically *Introduction to Data Mathematics*, that teaches linear algebra, data analytics, and R programming simultaneously using a project-based learning (PBL) approach. Then in DAR, students work in teams on open applied data analytics research problems provided by the clients. We describe the DAR organization which is inspired in part by agile software development practices. Students meet for coaching sessions with instructors multiple times a week and present to clients frequently. In a fully remote format during the pandemic, the students continued to be highly successful and engaged in COVID-19 research producing significant results as indicated by deployed online applications, refereed papers, and conference presentations. Formal evaluation shows that the pipeline of the single on-ramp course followed by DAR addressing real-world problems with societal benefits is highly effective at developing students' data analytics skills, advancing creative problem solvers who can work both independently and in teams, and attracting students to further studies and careers in data science.

**Keywords:** undergraduate education, data analytics, course-based undergraduate research, linear algebra, project-based learning, data visualization, machine learning.

# 1 Introduction and Goals for DAR

This paper reports on the course, *Data Analytics Research* (DAR), offered by the Mathematical Sciences Department at Rensselaer Polytechnic Institute (RPI). DAR is a course-based undergraduate research experience (CURE) in which undergraduate students conduct research on open-ended, real-world problems in applied data analytics.

Here we define applied data analytics (DA) to be the mathematics, models, and methods used to transform data into information to support data-informed decision-making. High-dimensional mathematical modeling is at the heart of DA, whether the methods used are from machine learning, statistics, image processing, or data visualization. Data-driven high-dimensional modeling and analysis complements physical mathematical modeling methods and is an essential and rapidly growing part of applied mathematics. Data-driven modeling is increasingly being considered as a fundamental component of applied mathematical education [13].

At RPI, we have developed a novel low-barrier pipeline through undergraduate mathematics to prepare students for careers in DA. Students complete one on-ramp foundational DA course that makes them eligible to enroll in DAR. In DAR, students work in teams to solve open-ended, real-world problems for clients using DA with coaching from the instructors. There are almost no lectures. The student teams are highly productive with their research, generating insightful presentations and reports for clients, refereed research papers, conference presentations, deployed software, and interactive applications. The combination of a low-barrier foundational DA on-ramp course followed by DAR has proven to be highly effective at recruiting diverse students into further study and careers in DA. Extensive computer science knowledge is not required, so the pool of students participating in DAR has been much more diverse in terms of gender and underrepresented minorities (URM) when compared to both RPI and the national computer science student populations from which machine learning DA students are usually drawn.

DAR, discussed here, is tightly integrated with the *Data Informatics Challenges in Technology Education (Data INCITE) Lab,* a joint research and education program run by Dr. Kristin Bennett, which focuses on providing applied research experiences in DA, interactive data visualizations, machine learning, and artificial intelligence with applications in diverse domains such as healthcare, manufacturing, business, and science. Data INCITE Lab has its own lab/classroom specifically designed to support team projects and CURE. DAR represents the institutionalization and expansion of the DA pipeline developed in part with funding from NSF Grant 1331023, *EXTREEMS-QED: Data Analytics Throughout Undergraduate Mathematics (DATUM)*[7] (2013-2017) and further funded by the United Health Foundation (2017-2020). To institutionalize these grants, the Data INCITE undergraduate research program evolved from a paid student internship to a primarily CURE model (i.e., for course credit) supplemented with a few paid student interns. DAR has helped the Data INCITE Lab to support over 250 DA research experiences for students since its inception.

The model of using an on-ramp or gateway course for the data science research pipeline is still relatively new among institutions of higher education. There are a few, like the University of California at Berkeley, that have established a DA pipeline, but not necessarily targeting mathematics undergraduate students. The 2018 National Academies of Sciences (NAS) consensus study report, *Data Science for Undergraduates: Opportunities and Options*, provides an overview of the types of data science-focused courses and programs targeted toward undergraduate students that are beginning to emerge [18]. NAS (2018) identified the following.

1. Integrated introductory courses that can satisfy a general education requirement;
2. A major in data science, including advanced skills, as the primary field of study;
3. A minor or track in data science, where intermediate skills are connected to the major field of study;
4. Two-year degrees and certificates;
5. Other certificates, often requiring fewer courses than a major but more than a minor;
6. Massive open online courses, which can engage large numbers of students at a variety of levels; and
7. Summer programs and boot camps, which can serve to supplement academic or on-the-job training.

The learning outcomes identified for students who successfully complete DAR are the ability to:

1. Model real-world observational datasets with an informed appreciation of data life cycles and underlying processes;
2. Ask key questions to find insights into the model and data in order to create knowledge from data;
3. Apply DA tools to implement, analyze, and visualize models and data;
4. Understand the expanding role of DA in fields such as healthcare through greater appreciation of relevant datasets, problems, and data-driven solutions; and
5. Communicate results effectively to diverse audiences in visual, written, and oral formats.

DAR is an example of the effective use of project-based learning (PBL) strategies to afford students the opportunity to develop DA knowledge and skills in context. PBL also encourages students to become creative mathematicians by asking their own questions and developing their own strategies for addressing these questions. "PBL often begins with a project to motivate and support the construction of knowledge through active learning with all members of the learning community, teacher and learners, engaging in the problem solving and questioning process"[20]. DAR focuses on coaching students in teams through open research problems and the assignment, course, and evaluations are designed to engage and challenge students to become effective, creative problem solvers and self-directed learners. There are no exams and very few lectures. This DAR course framework translated very well to a fully remote format during the COVID-19 pandemic.

This paper highlights the student research projects in DAR, describes the DA pipeline and on-ramp courses, expounds the PBL-based DAR framework, details the assessment strategies that continually provide both formative and summative feedback, summarizes the diversity evaluation and outcomes evaluation results, reviews the impact of the PBL framework and DA pipeline for undergraduate math research experiences, illuminates potential pathways to establish similar DARs, and shares lessons learned including both benefits and challenges as part of the discussion.

# 2 DAR Research Projects

## 2.1 Research Projects Overview

In DAR, teams of four to six students from a total course enrollment of 20-30 students work on open applied DA research questions provided by real-life industry, university, and community clients. The research problems are drawn from the research activities and/or contacts of the Data INCITE Lab. The instructors review potential relevant data and work with clients, and then pre-select and define the research problems in advance of the class. Usually, the data sources are at least partially identified. We typically start with a high-level research problem, such as "What are the social determinants of COVID-19 mortality?" Most often, the research question has *not* been transformed into a modeling and analysis workflow before the DAR. The instructors frequently use DAR to initiate work in new research areas of interest or to more widely explore questions brought up in existing funded research contracts. Some projects may be incubator projects which are entirely new projects. Most are ongoing projects that address additional research questions designed to complement prior DAR or funded research projects.

In DAR, projects are presented by the clients or instructors at the start of the term; students submit an application indicating their project preferences, their skills, and prior coursework and research. The instructors construct teams that attempt to balance students' interests and experiences. The individual teams might work on completely distinct research topics or on different aspects of a single project. Often, DAR research leads to significant research outcomes but they may not be completed in a single semester. Typically, a combination of undergraduate student research teams, engaged in course and internship experiences over multiple semesters and summers, bring a project from inception to published papers and deployed applications; DAR is the core of this pipeline.

To ensure the success of the Data INCITE Lab initiatives, we usually use a small team of undergraduate students to do work in advance to refine the problem and define the data sources. A team of one to four students will often work as paid or for-course-credit interns to perform the initial data preparation, exploration, and analysis to set up the project for the course. They identify data sources, prepare and clean the data, do preliminary data exploration, data visualization and investigations to help refine and focus on the research questions. R notebooks are used to document these preliminary results and the datasets being used. These are placed in project-specific GitHub repositories for use in DAR. Typically, one or more students from the data preparation "advance team" will also be enrolled in the course and serve as team leads. Potentially multiple teams will then work on the project in DAR itself. For large problems, the original problem may be broken into subtasks. For example, the entire Summer 2019 class worked on MORTALITYMINDER, but we divided the students into four teams each with distinct responsibilities, e.g., data preparation and handling missing data, developing modeling and analysis of social determinants, user interface design, and backend design of the web application.

For some projects, the focus is on DA research leading to publication, while for others the objective is to produce an interactive, analytics-driven application. In both cases, the bulk of the work will be completed during DAR, with a small team of students engaged during the subsequent term to finalize their team's work for publication or, in the case of a web app, to add any missing functionality and polish the interface.

Between May 2019 and December 2020, DAR students worked on health informatics related research with thanks to generous support from the United Health Foundation. All DAR projects were created in R with interactive applications built using R Shiny[9], except for *COVID Twitter* which was based on Python. A gallery of projects created in DAR can be found at `https://idea.rpi.edu/research/projects/data-incite`. We briefly list the projects here and then discuss some projects in more detail. Clients are noted for each. To simplify referring to the assets produced by the students, we provide links and references to outputs which include online apps (A), software in GitHub (G), refereed papers (R), manuscripts (M), and presentations (P) at conferences or workshops by students or Dr. Bennett. The COVID-19 related apps are also described in a video [5]. Sample snapshots from the apps and research results are shown in Figures 1-4.

1. MORTALITYMINDER: Web-based visualization tool that enables interactive exploration of social, economic, and geographic factors associated with premature mortality in the United States that won 3rd place in 2020 United States Health and Human Services (HHS) Agency for Healthcare Research and Quality (AHRQ) Visualization Resources of Community-Level Social Determinants of Health Challenge. (Clients: Health Care Executives) A=[23], G=[22], M=[8], P=[6, 4, 14].

2. COVIDMINDER: App analyzing regional disparities in COVID-19 cases, mortality rates and social determinants of COVID-19 pandemic. (Clients: RPI Researchers) A=[30], G=[29].

3. COVID Back-to-School: App for generating actionable information on how to reopen schools based on past COVID-19 infection rates, predictive modeling and selected social distancing strategies. (Clients: RPI Researchers) A=[24], G=[27].

4. Social Determinants Associated with COVID-19 Mortality: Analysis of socio-economic factors associated with increased COVID-19 Mortality. (Clients: RPI and External Researchers) RP=[11, 12, 37], G=[34].

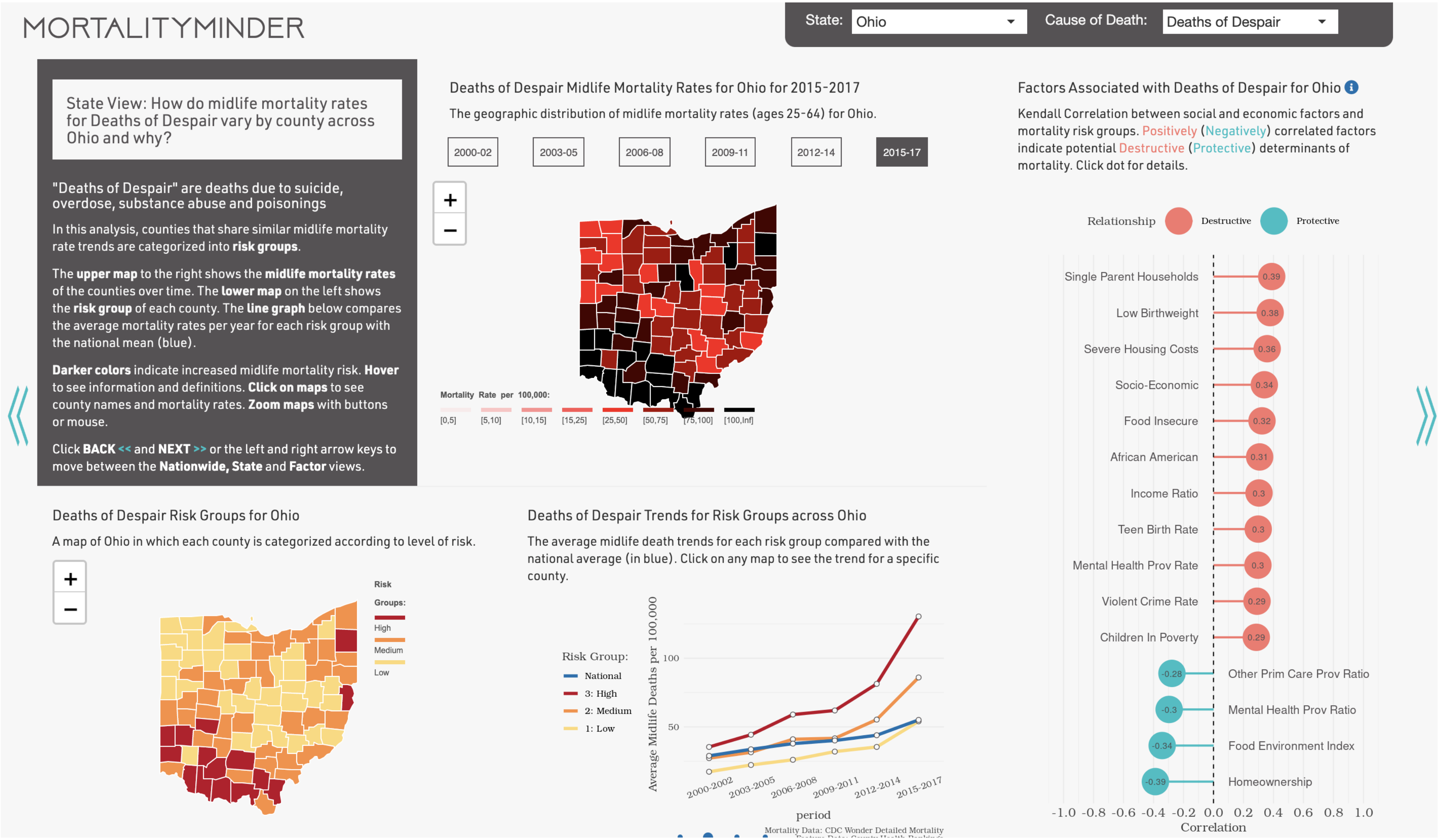
5. COVID WarRoom: App for predicting COVID-19 infection rates at a location based on past COVID-19 infection rates, predictive modeling and selected social distancing strategies. (Clients: RPI Researchers) A=[28], G=[27].

6. COVID Twitter: Natural Language Processing (NLP) tool for analysis of COVID-19 Twitter discourse used to study mask wearing sentiments. (Clients: RPI Researchers) G=[26, 25], RP=[19], P=[17].

7. RPI SafeCampus: App to visualize and understand congestion at RPI using anonymous WiFi access point data. (Clients: RPI Researchers) A=[33], G=[31], P=[10].

8. RPI StudySafe: App that helps students find a safe place to study based on WiFi access point usage. (Clients: RPI Researchers) A=[35], G=[32].

9. Analysis of Circadian Rhythms in Omics Data: Analysis and Visualization of Circadian Proteomic, Transcriptomic, and Pathway Data. (Clients: RPI Researchers), AG=[15], M=[16]



MORTALITYMINDER

State: Ohio

Cause of Death: Deaths of Despair

State View: How do midlife mortality rates for Deaths of Despair vary by county across Ohio and why?

"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

In this analysis, counties that share similar midlife mortality rate trends are categorized into **risk groups**.

The **upper map** to the right shows the **midlife mortality rates** of the counties over time. The **lower map** on the left shows the **risk group** of each county. The **line graph** below compares the average mortality rates per year for each risk group with the national mean (blue).

**Darker colors** indicate increased midlife mortality risk. **Hover** to see information and definitions. **Click on maps** to see county names and mortality rates. **Zoom maps** with buttons or mouse.

Click **BACK <<** and **NEXT >>** or the left and right arrow keys to move between the **Nationwide, State** and **Factor** views.

Deaths of Despair Midlife Mortality Rates for Ohio for 2015-2017

The geographic distribution of midlife mortality rates (ages 25-64) for Ohio.

Factors Associated with Deaths of Despair for Ohio

Kendall Correlation between social and economic factors and mortality risk groups. Positively (Negatively) correlated factors indicate potential Destructive (Protective) determinants of mortality. Click dot for details.

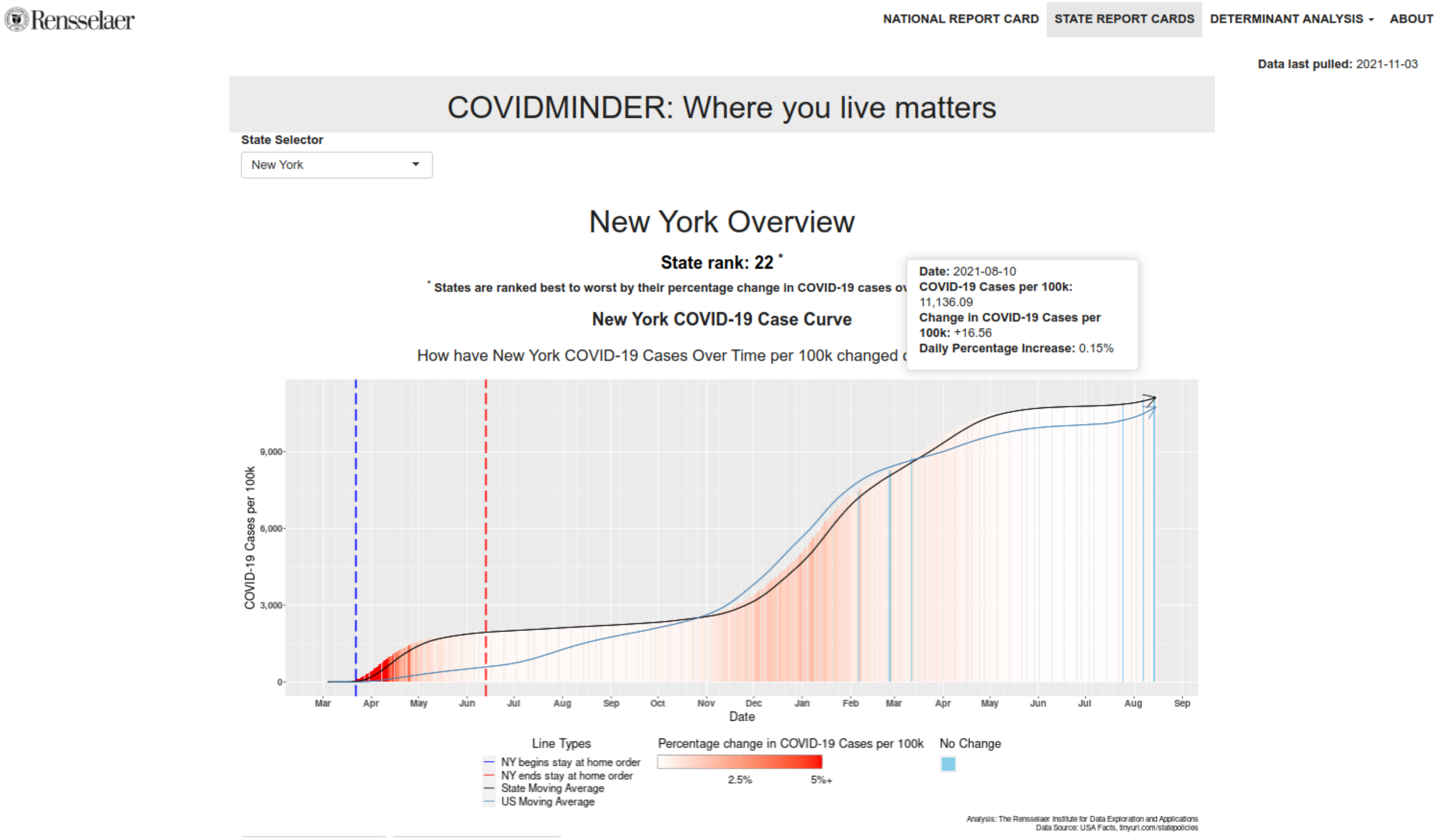
Deaths of Despair Risk Groups for Ohio

A map of Ohio in which each county is categorized according to level of risk.

Deaths of Despair Trends for Risk Groups across Ohio

The average midlife death trends for each risk group compared with the national average (in blue). Click on any map to see the trend for a specific county.

Figure 1: Screenshot from a DAR App implemented in R Shiny: MORTALITYMINDER



**Data last pulled:** 2021-11-03

COVIDMINDER: Where you live matters

**State Selector**

New York ▾

New York Overview

**State rank: 22** *

* **States are ranked best to worst by their percentage change in COVID-19 cases ov**

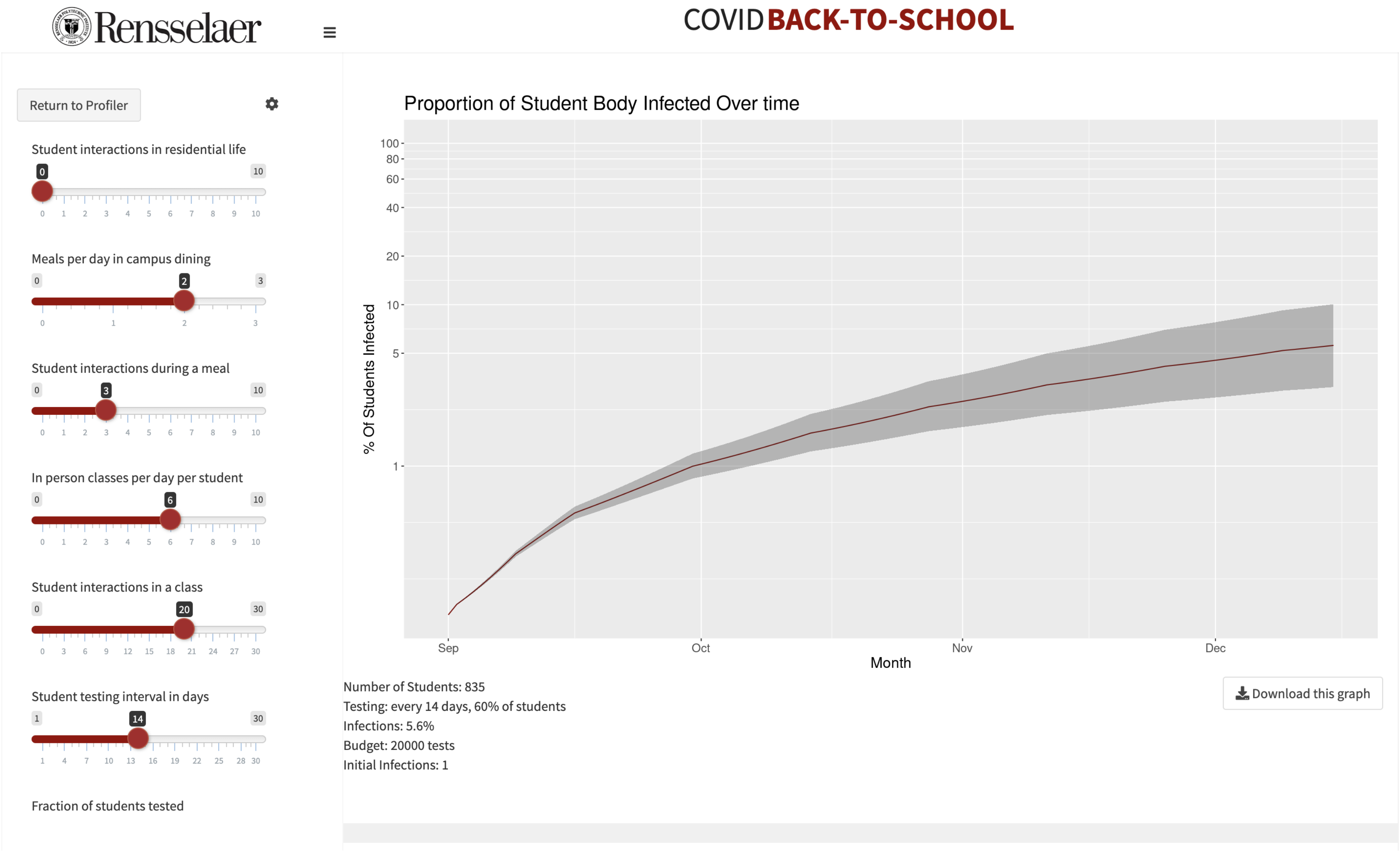
**New York COVID-19 Case Curve**

How have New York COVID-19 Cases Over Time per 100k changed c

**Date:** 2021-08-10
**COVID-19 Cases per 100k:** 11,136.09
**Change in COVID-19 Cases per 100k:** +16.56
**Daily Percentage Increase:** 0.15%

Analysis: The Rensselaer Institute for Data Exploration and Applications
Data Source: USA Facts, tinyurl.com/statepolicies

Figure 2: Screenshot from a DAR App implemented in R Shiny: COVIDMINDER



Student interactions in residential life: 0

Meals per day in campus dining: 2

Student interactions during a meal: 3

In person classes per day per student: 6

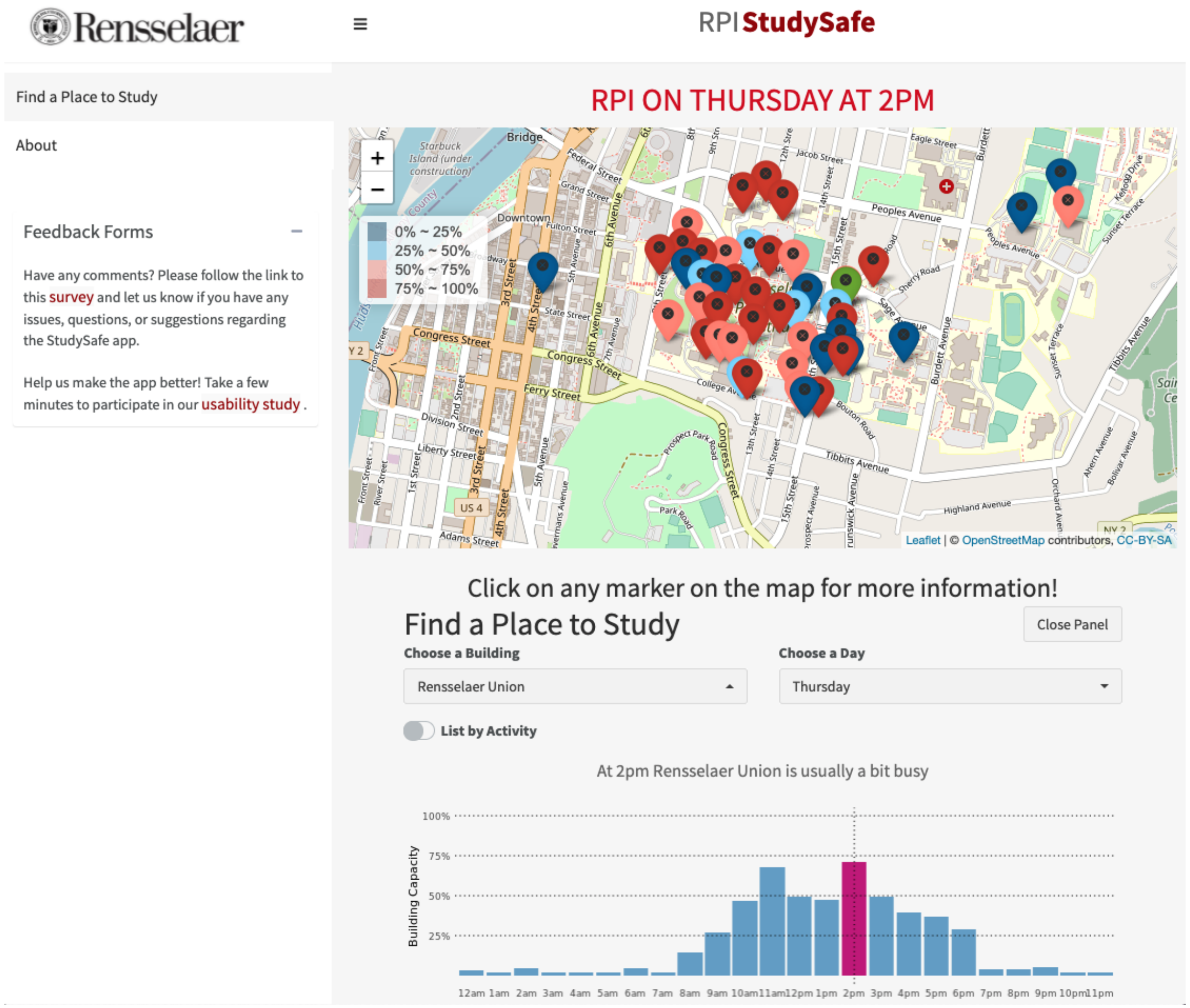
Student interactions in a class: 20

Student testing interval in days: 14

Fraction of students tested

Number of Students: 835
Testing: every 14 days, 60% of students
Infections: 5.6%
Budget: 20000 tests
Initial Infections: 1

Figure 3: Screenshot from a DAR App implemented in R Shiny: COVID Back-to-School



Have any comments? Please follow the link to this survey and let us know if you have any issues, questions, or suggestions regarding the StudySafe app.

Help us make the app better! Take a few minutes to participate in our usability study .

Click on any marker on the map for more information!

Find a Place to Study

Choose a Building: Rensselaer Union

Choose a Day: Thursday

List by Activity

At 2pm Rensselaer Union is usually a bit busy

Figure 4: Screenshot from a DAR App implemented in R Shiny: RPI StudySafe

## 2.2 Sample Web Application Projects: MORTALITYMINDER and COVIDMINDER

The goal of MORTALITYMINDER is to give healthcare providers (e.g., hospitals), payers (e.g., insurance companies), researchers (e.g., investigators), and policymakers (e.g., state government) actionable insights into how, where, and why midlife mortality rates are rising in the United States. The MORTALITYMINDER app was created as an entry in the 2020 United States Health and Human Services (HHS) Agency for Healthcare Research and Quality (AHRQ) *Visualization Resources of Community Health-Level Social Determinants of Health Challenge*. The app won third place overall and was the only academic team to win an award [1]. Shown in Figure 1, MORTALITYMINDER is a web-based visualization tool that enables interactive exploration of social, economic and geographic factors associated with premature mortality among midlife adults ages 25-64 across the United States. Using authoritative mortality data from the CDC and social and economic data primarily from County Health Rankings `https://www.countyhealthrankings.org/`, MORTALITYMINDER uses clustering, statistical analysis, and data visualization in an interactive app. It is designed to help healthcare payers, providers, and policymakers at the national, state, county and community levels identify and address unmet healthcare needs, healthcare costs, and healthcare utilization. MORTALITYMINDER is an open source project implemented on the R Shiny platform.

The MORTALITYMINDER project followed our now-standard project model. Prior to the now official DAR course, a team of six students worked as paid interns for six weeks in Summer 2019 to do initial data preparation and modeling. Then, with help from the members of this pre-course data preparation team, the 22 additional students did the primary exploration and analysis for the project during a compressed six-week summer course after which the app was finished by a mixed team of veterans and new students using the paid internship model the following semester. The teams of students focused on different aspects of the problem and worked on reviewing the state visualization and analysis, nationwide analysis and visualization, analytics (including missing data imputation), and infrastructure (including data preparation and R Shiny app development). The teams presented their work approximately weekly and received feedback from our clients, an *ad hoc* advisory board of three vice presidents from health care companies and a senior manager from the New York State Department of Health. MORTALITYMINDER was presented at the American Medical Informatics Association (AMIA) Annual Symposium 2020 and a regional meeting [6, 14].

MORTALITYMINDER served as inspiration for COVIDMINDER, a web-based visualization tool that uses publicly available data from multiple sources and interactive analysis to reveal the regional disparities in outcomes, determinants, and mitigations of the COVID-19 pandemic (see Figure 2 (b)). COVIDMINDER was initiated by a team of students in early 2020 on six-week internships and continued as a six-week DAR during the Summer 2020 term and as a 16-week DAR during the Fall 2020 term.

## 2.3 Sample Research Project: Social Determinants of COVID-19 Mortality

MORTALITYMINDER's analytical social determinants data and COVIDMINDER's COVID-19 mortality data served as foundations for DAR student research in social determinants associated with COVID-19 mortality. Students identified which social determinants were associated with significant changes in COVID-19 mortality rate at the county level as of July 5, 2020. Uniquely, we accounted for a comprehensive list of comorbidities and the impact of differing state policies such as closings and reopenings. An association study was conducted using the significant high-risk factors as controls to evaluate 41 social determinants in order to find those associated with COVID-19 mortality. The association study utilized negative binomial mixed models to analyze county level data (n=3093 counties), statistically corrected for possible false discoveries using the Benjamini-Hochberg Procedure. Model performance and fit were validated using a procedure from Wu et al [36]. The undergraduate student research work was accepted for poster presentation at AMIA 2020 [12]. Further details on the data sources and analysis are described in the full manuscript. Motivated by the possibility of seeing their work carried through to publication, an undergraduate from DAR continued the work, creating an expanded version of the paper examining how social determinants associated with COVID-19 mortality changed over time. Using US county-level data from July 5 and December 28, 2020, the effect of 19 high-risk factors on COVID-19 mortality rate were quantified and association studies for 40 social determinants were performed yielding valuable insights into how social determinants of COVID-19 mortality evolved during the pandemic. This research was accepted and presented as a full paper during the 2021 AMIA Annual Symposium [11].

# 3 Data INCITE Pipeline

The Data INCITE Pipeline for mathematics students consists of an on-ramp foundational course such as *Introduction to Data Mathematics (IDM)* followed by a research experience, now DAR. First established with the NSF DATUM grant in 2013, DATUM demonstrated that DA literacy can be incorporated effectively into existing mathematics curricula via the addition of a few key foundational courses and undergraduate research experiences in applied DA. The innovative IDM, developed at RPI, transformed the traditional math curriculum to teach students high-dimensional DA and modeling at an early stage in their academic careers.

The goals for DAR are anchored by the Data INCITE Pipeline program goals, designed to address the growing need for a cadre of individuals willing and able to engage in DA. The Data INCITE Pipeline goals include:

1. Developing student skills in DA;
2. Recruiting students early in their undergraduate studies, affording them increased options to pursue further studies and careers in data science;
3. Creating low-barrier educational pathways through mathematics for DA, distinct from the more traditional pathways that exist in computer science or statistics;
4. Increasing diversity of students pursuing careers in DA, both in terms of traditionally underrepresented minorities and genders and in terms of students from a more diverse set of majors including mathematics, pre-med, biology, biomedical engineering, and computer science; and
5. Recruiting students to pursue careers in DA by engaging undergraduates in research through internships and CUREs that use project-based learning (PBL) to solve real-world problems.

Students can enter the Data INCITE Pipeline as early as their first or second year of undergraduate study by taking IDM or an alternative on-ramp foundational DA course. The pipeline is low barrier in that only lower level math courses and one computer science course are required. In IDM, students learn the basic linear algebra and DA techniques which are sufficient for them to fully engage in all phases of the data life cycle — problem definition, preparation, exploration, modeling, interpretation, and communication — in a PBL environment. IDM simultaneously teaches students basic linear algebra and powerful DA techniques. The philosophy of IDM is to provide early, focused, and deep mathematical and DA experiences that enable students to engage in creative problem solving on open, compelling real-life problems using high-dimensional mathematics and computation as part of a community of learners. The program goes deeply into a focused set of linear algebra, geometry, and DA subjects. IDM incorporates a set of projects in which students analyze data using one or more of the techniques developed in the course including classification, principal component analysis (PCA), data visualization, and clustering. Components of the mathematics behind the techniques are presented as well. As created in 2014 by Dr. Kristin Bennett and Dr. Bruce Piper, both professors in the department of mathematics and co-instructors for the initial course offerings, IDM originally used MATLAB. In 2017, we switched to R since it is a powerful open source language for DA that is widely used in industry and research. IDM prepares students to do creative practical DA problem solving on open real-life projects.

The idea is that by engaging students in DA early in the math curriculum, students are more motivated to learn subsequent topics as they arise and to persist in the field of mathematics. Alternative on-ramp courses, such as Biostatistics, Data Analytics, and Machine Learning, have now been added. In these courses, students learn higher-dimensional mathematics, statistics, or machine learning within the context of DA problems, enabling the students to join in applied DA research projects for real-life clients. All the fundamental mathematics topics covered in the on-ramp courses may also be covered in other courses, albeit frequently at a more abstract level, and students' learning is further deepened by subsequent exposures. Importantly, students learn DA which will be quite valuable throughout their undergraduate career and beyond. The combination of the gateway foundational course plus an early applied DA research experience was very successful at both developing students' DA skills and recruiting students for careers in DA [20]. As a testament, many students do elect to enter careers in DA and/or pursue graduate work related to DA.

Originally, the DA paid summer research internships were offered within the Data INCITE Lab as part of the pipeline established with the DATUM grant, and a CURE DA senior student capstone course was offered. To make these research opportunities scalable and sustainable for students early in their undergraduate studies, the DA summer research internships were gradually transformed starting in 2017 into credit-bearing CUREs in DA research, now called DAR. As a credit-bearing course with only one pre-requisite on-ramp foundational course in DA, DAR enables many more students to engage in research.

At RPI, the majority of undergraduate students in mathematics continue to do research within DAR with some additional students still hired as student research interns. For many students, IDM and DAR are their first experiences with DA and research. Following only these initial experiences, 85% of the students indicated that they were interested in pursuing future work in DA. Students' participation in DAR research projects continues to be very successful and has resulted in published papers, conference presentations, and applications including the prize-winning MORTALITYMINDER application. With the support of the United Health Foundation from 2018-2020, DAR and the Data INCITE Lab focused on health informatics research problems using DA. The added emphasis on the health context in DA proved to be particularly compelling to students as the Data INCITE Pipeline pivoted to primarily focus on using DA research to understand and respond to COVID-19 challenges once the pandemic began.

# 4 Project-Based Learning (PBL) DAR Framework

Engaging students in DA research early in their undergraduate studies through real-life projects has proved very successful [20]. The PBL-infused DAR framework, including the instruction infrastructure and organization of the CURE, is described in this section.

## 4.1 DAR Instructional Infrastructure

Almost all DAR projects are done in R using R Studio except for a few minor exceptions when students are allowed to use Python. Student research is created and documented in R Markdown notebooks. The technical infrastructure for the course continues to be hosted by the Rensselaer Institute for Data Exploration and Applications (IDEA), which made shared access to R Studio Server and storage available on its high-performance compute cluster. This ensured that every student in the course had an identical compute environment and access to shared files, eliminating the problems inherent with students having diverse, potentially poor-performing personal computers, operating systems, and other computational tools required for the course. This same environment is used in the on-ramp IDM course.

DAR adopted collaboration tools commonly used across the software development community, including GitHub for source code management and issues tracking and Slack for intra- and inter-team chat-based communications [2, 3]. The use of GitHub allowed individual teams to collaborate and contribute across time and distance on complex projects without concerns that they might damage their teammates' work, and with confidence that harmful code could be reverted and new approaches taken. GitHub also enabled the instructional team to more readily oversee and manage work across the course. Slack was used for both project communications and course administration. With members of the instructional team attentive and responsive to notifications, questions and issues posed by students could be dealt with often in real time, eliminating barriers to their progress.

DAR is a team effort of Dr. Kristin P. Bennett and Dr. John S. Erickson. Dr. Bennett is the founder and director of the Data INCITE Lab, and Professor of Mathematical Sciences and Computer Science at RPI. As associate director of the Rensselaer IDEA and lead researcher on many data related research projects, Dr. Bennett is responsible for the research directions of the lab and provides the mathematical modeling and deep analytics coaching. Dr. Erickson is the Director of Research Operations for the Rensselaer IDEA. He works extensively with students in the Data INCITE Lab, including supervising undergraduate research projects and team meetings, R and GitHub coaching, and providing instructional technology and infrastructure support. Drs. Bennett and Erickson together teach one CURE course containing 20 to 30 students at a time. This contrast with Rensselear's Undergraduate Research Program model in which a professor supervised the work of one or two students for credit at a time.

The team leadership approach that combines extensive experience in data science research, technology infrastructure, and project management enables DAR to provide highly engaging undergraduate research experiences within the course while simultaneously producing significant research results for real-life problems. Having at least one instructor knowledgeable about data analytics research seems essential for DAR. Dr. Bennett did run prior senior capstone CUREs and internship programs with 10-20 students alone, but they did not produce the publishable results and interactive applications now routinely created in the team taught course. Both instructors spend most of the class period and significant outside time interacting with students in team meetings, Slack, and office hours. But the total time devoted to class and class preparation is not more than that required for a typical class because there are few lectures. Grant support enabled the development of the present team-taught coaching model for DAR, provided for the computational needs of the course, and provided student interns who prepare, continue, and complete projects and applications when the DAR course is not in session.

## 4.2 DAR Organization and Processes

DAR has the following organization.

1. We start by having real-life university, industry and community clients (possibly one of the DAR instructors) present the problems. These clients are recruited through the instructors' personal and professional networking, within the university, the research community, and RPI alumni. We also frequently recruit clients and projects as spin-offs from existing research grants. The students then fill out a preference form to indicate which problems or sub-problems they prefer, why they want to work on that problem, their majors/minors, and any prior coursework and experiences. The students are formed into the teams that they will work with for the remainder of the semester. We generally honor student preferences, but also work to build balanced teams based on students' prior experiences and interests.

2. Usually one or more students from the pre-term data preparation team continue in DAR and act as project mentors and role models for other students. Even though the advance team students have worked with the project data for a time before the semester, they generally have completed similar levels of coursework (e.g., perhaps only IDM) as DAR students, so they are are just a few weeks ahead of their classmates. Students who were part of the data preparation team are typically high-achievers who have been selected by application on a competitive basis, whereas all students completing an on-ramp DA course are permitted to register for DAR.

3. The primary assignments are implemented through R notebooks, which serve multiple purposes. They encourage students to make steady progress, to document their work for assessment, to capture work so it can be shared across the team, and to help students establish good research habits. There are no right or wrong answers in the R notebooks; every notebook is different because every student's research path is different. Students are encouraged to collaborate with other students and to reuse code, but they must document their individual contributions as well as their collaborators. The *R Project Notebook Rubric* that we use allows for students to pursue their own unique analysis paths. No specific mathematics or DA is required. The rubric addresses the following elements.

   (a) Work Summary Notebook committed to GitHub

   (b) Weekly Work Summary

   (c) Student Contribution: Clearly defined, unique contribution done by the student (i.e., analysis, code, ideas, writing)

   (d) Discussion of Primary Findings

   (e) Code Quality

   (f) Writing Quality

   (g) Scientific Clarity and Reproducibility

   (h) Extra credit for Creativity and Significant Project Insights

4. These first assignments in DAR require the students to tackle an open-ended problem within the real-life project and to document their project tasks and findings. Starting with an R notebook in GitHub, we ask students to explore a question, update the notebook with their results committing it to GitHub to submit, and be prepared to share their notebook and discuss results during the next

meeting. After that, the students' notebooks are used to document their analyses and projects. Students submit their notebooks roughly bi-weekly in the full-semester DAR, and weekly in a 6-week DAR. Their final notebooks summarize the students' most significant contributions to the their respective projects in DAR.

5. The second type of assignment centers on team presentations. Each team presents regularly to the clients and instructors. The culmination of DAR includes a mini-conference in which each team presents its work to the clients and an invited audience. The entire class attends all team presentations to learn from the activities of other teams.

6. As a PBL-infused research experience, instructors primarily interact with the students through coaching sessions. After a few preliminary lectures to get students started, there is a shift with the remainder of the course consisting of meetings between the teams, instructors, and clients. The core principles of agile software development were adopted, including: a focus on teamwork and communication; an emphasis on producing and sharing working code; collaborating closely and communicating often with clients; and extreme flexibility (i.e., responding to change over following a plan as the situation demands). "Agile" provides the basis for interactions between and among students, the instructors, and our clients. The instructors meet with each team at least twice a week. During the first meeting each week, a team will have a "stand up" session designed for quick problems. As part of the "stand up", students quickly present what they did, any problems they encountered, and what they plan to do next. The second meeting of each week consists of a "deep dive" meeting with the expectation that students will present and/or demonstrate progress based on their notebooks. Thus, students gain extensive experience presenting their work informally and getting feedback before the more formal team presentation. We normally require closely related project teams to meet together to both help coordinate projects that involve multiple teams and to expose students to more analysis experiences. It may be that one team will alternate with the other team in doing the stand up and the deep dive.

7. Since every student works on different mathematical and DA methods within a project, students need to become self-directed learners. To assist students, some of the class time is approached as office hours with additional access to office hours outside of class available to students as well. R Bootcamps and specialized R tutorials are held as needed, which complement the Data INCITE Lab goals and many other RPI courses. We also hold *RPIrates*, a monthly student-centered seminar series focused on R skills development. The R Bootcamps, tutorials and RPIrates meetings are conducted by Dr. Erickson. They are held outside of class-hours and also made available to students via recordings.

## 4.3 DAR Coaching During Remote Learning

The coaching model provided both DAR students and DAR paid student interns with very engaging remote experiences. Instructors met with students in teams at least twice per week, asking them individually in each session to summarize *what did you do*, *what did it mean*, *what are you going to do next*, and *what obstacles are you facing*? Students gave regular team presentations, met with real-life clients, received individual feedback, and created results and apps that were real contributions to understanding the spread of COVID-19 in the United States. During past in-person CURE offerings, some students could partially hide behind the productivity of other team members' work. In the remote model with regular meetings and required online activity, individual student contributions are much clearer. Virtual meetings put students more "on the spot" as individuals while compelling them to integrate better as a team. This mode of operation can be a bit overwhelming and intimidating at first for some students. We provided a lot of individual coaching via additional office hours to help individual members contribute to the best of their ability. We use the same coaching model and team meetings with students enrolled in the DAR and the DAR student interns.

DAR was conducted both in-person and remotely. In many ways, remote instruction increased our productivity because the tools used for remote learning are the same ones that make for highly productive team research and work. We embraced videoconferencing when interacting with clients; utilized R packages such as R Shiny for web delivery of DA app work-in-progress; and leveraged best-in-class online collaboration tools for team project working including as GitHub, Slack, and Overleaf. Students also rapidly embraced presenting their work online every meeting, frequently asking to share their screens to show off their most recent accomplishments. To keep students engaged, we scheduled a mandatory remote team meeting for each team outside of class hours. We used a flipped-classroom with recordings for R programming and other technologies used in outside sessions made available to students so that we could devote the full class time to engaging with individual students, meeting with each team in turn. The approaches and techniques we learned transitioning to remote learning during COVID-19 continues to be used in DAR when we return to the in-person classroom setting. The in-person DAR offered in fall 2021 continues to have remote outside-of-class team meetings, remote weekly client meetings, and flipped lectures.

## 4.4 DAR and Design Features of the PBL Framework

Students interact with the clients throughout DAR and ultimately prepare their findings and presentations for the real-life clients providing students with a complete and authentic DA learning research experience.

The features of the Data INCITE Lab and DAR are based on the four design principles for STEM-based PBL proposed by Slough and Milam (2013)[21]:

1. making content accessible;

2. making thinking visible, which includes using visual elements to help the learner and using learner constructed visual elements to assess learning;

3. helping students learn from others;

4. promoting autonomy and lifelong learning.

As discussed in Seddon 2021[20], the proposed four design principles can be pedagogically conceptualized as they are in DAR to include:

1. using real-world content and problem solving in context for greater accessibility;

2. using data visualization tools to support problem solving and communication;

3. incorporating interaction and collaboration among all in the “community of learners” to enable learning from each other and creating “new experts” in the field; and

4. supporting increases in knowledge, confidence and beliefs as part of promoting ongoing interest in future learning and use of data analytics [20].

These four revised design principles served to guide the design for the PBL-infused CURE and reflected the prior DATUM undergraduate research program design principles in the context of DA. Seddon’s evaluation of DATUM found that:

> Making content accessible was supported through the use of real-world projects that students could relate to and interact with...The real-world projects render the material more accessible and, arguably, provide a “cognitive hook” that facilitates recall of the learning, supporting future construction of knowledge. Thinking was made visible through the use of data visualization technologies and strategies incorporated into the data analytics [undergraduate] course work and ... research experience. Students learned from each other through interactions and collaborations that were integral to their projects. Autonomy and lifelong learning were promoted through the building of knowledge, confidence and beliefs that increased opportunity and success for future plans involving data analytics. All of the learning and research activities were further supported in the environment encompassed by the teacher, learner and the project. [20]

Table 1: Survey Respondents by Major, Gender, Ethnicity

| **Characteristic** | **Response** | % |
|---|---|---|
| Major (n=118) | Mathematics | 29.7 |
| | Computer Science | 16.1 |
| | Other | 13.6 |
| | Biology | 11.0 |
| | Engineering | 9.3 |
| | Biochemisty and Biophysics | 7.6 |
| | Management | 0.8 |
| | None given | 11.9 |
| Gender (n=118) | Female | 51.0 |
| | Male | 49.0 |
| Ethnicity (n=77) | White | 47.0 |
| | Asian | 27.0 |
| | Hispanic | 16.0 |
| | Black or African American | 6.5 |
| | Multiple or Other Races | 3.5 |

# 5 DAR Assessment and Evaluation

We report here on the combined assessment during 2020 of the spring term IDM course (84 students enrolled for course credit), the summer term DAR (14 students enrolled for course credit + 14 paid student interns) and of the fall term DAR (24 students enrolled for course credit + 5 paid student interns). We combined the analysis for all the students in 2020 to give an indication of the overall effectiveness of the pipeline beginning with IDM followed by DAR courses and to provide more power to the analysis. We used a student survey designed to assess the core learning outcomes discussed in the introduction section of this paper. The survey instruments were originally designed for pre-experience evaluation and post-experience evaluation as part of the DATUM program at RPI. The survey instruments were converted for use as a single post-experience evaluation for use in DAR to evaluate student perspectives and plans in major areas: learning outcomes, career trajectory, and diversity. Cronbach's analyses of the survey instruments completed for the DATUM program confirmed internal reliability for the survey item constructs that included knowledge, ability, gains, confidence, beliefs, and future plans [20].

## 5.1 Diversity Evaluation

A total of 118 responded to end-of-term surveys for the two DAR course offerings and for IDM during 2020 (see Table 1). 64.6% of these students were drawn from four majors including math (30%), computer science (16%), biology (11%) or biochemistry/biophysics (7.6%), with the remaining students from other or unspecified majors. Of the 118 respondents, there were more females (51%) than males (49%), which marks a significant overrepresentation of females relative to the RPI undergraduate population of female (32%) and male (68%). This is a remarkable achievement since data science programs that target science and engineering are typically less diverse than the undergraduate population in terms of gender. In the United States in 2017-2018, only about 20% of undergraduate engineering and computer science degrees were granted to females.[1] This may reflect that we include students from a diverse set of majors; nationally, women make up almost 60% of biology majors and almost 40% of math majors. In terms of ethnicity for the 77 respondents for whom we have data, 47% were White, 27% were Asian, 16% were Hispanic, and 6.5% were Black or African American; 3.5% were multiple or other races. This indicates an overrepresentation with respect to the RPI undergraduate population that is Hispanic (9.7%) and Black or African Americans (4.1%). Our focus on low-barrier pathways into DA and/or data science through DAR enabled students with only one computer science class to enroll, and it expanded both the number of students and the diversity of students in the Data INCITE Pipeline.

[1] https://research.swe.org/2016/08/degree-attainment/

Table 2: Survey Responses Related to Perceived Impact on Data Analytics Skills. Percentages indicate responses to “To a great extent” or “Somewhat” combined. (Additional response options not shown included “Very little” and “Not much at all”.)

| **Question “Please indicate to what extent your experience in this class (including lessons, assignments, projects, etc.) contributed to your ability in each of the following areas?”** | **“To a great extent” or “Somewhat” (%)** |
|---|---|
| Working with data sets | 83 |
| Knowledge of existing applications (e.g., R) to use data to support problem solving, knowledge gathering, and/or decision making | 84 |
| Selecting an appropriate mathematical or statistical model to analyze data | 80 |
| Preparing data for analysis or modeling | 81 |
| Using computation to solve a data analysis or modeling problems | 75 |
| Knowledge of the mathematics underlying data modeling or data analysis | 72 |
| Using statistical models for data analysis | 78 |
| Identifying trends, drawing conclusions from data | 81 |
| Predicting future outcomes using data | 76 |
| Interpreting results from data analysis | 80 |
| Validating the quality of a result from data analysis or modeling | 72 |
| Determining if the assumptions made in modeling data or analyzing data are valid | 70 |
| Communicating the findings resulting from a data analysis or modeling to a broad audience (written or verbal presentations) | 71 |
| Working in interdisciplinary teams to solve data-based problems | 64 |

## 5.2 Outcomes Evaluation

The students’ responses to questions relevant to the learning outcomes are reported in Table 2. On average, 76% of students responded “To a great extent” or “Somewhat” when asked 14 questions related to how they perceived the extent that the DAR or IDM experience contributed to DA knowledge and skills. The skills with the greatest percentage of respondents indicating the DAR experience contributed “To a great extent” or “Somewhat” included using R to support data-driven problem solving (84%), working with datasets (83%), preparing data for analysis (81%), identifying trends in data (81%), and interpreting results (80%). This survey instrument was developed to assess all DATA INCITE Lab programs in the DATUM and UHF grants. One class cannot address all DA skills, and thus we do not expect students to achieve significant improvement in all of the outcomes in a single course. In DAR, students’ individual research typically emphasizes different DA skills depending on their project and their roles on the team. In addition, since these are courses, students applied different levels of effort and thus achieved different levels of mastery of DA skills. This was especially true for IDM during 2020 when many students took advantage of liberal pass/fail policies during the pandemic.

The Data INCITE Pipeline was highly effective in recruiting students to further studies and careers related to DA. As reported in Table 3, the majority of students in IDM and DAR (on average 76%, with a range of 62% to 85%) responded “Strongly agree”, “Somewhat agree” or “Agree” for question items related to their future plans and perceived value of DA in their academic and career paths. The question items with higher percentages of agreement follow.

- 85% of students were planning to pursue additional courses and experiences in DA.

Table 3: Survey Responses Related to Future Plans and Perspectives related to Academic and Career Path. Percentages indicate responses to "Strongly agree", "Somewhat agree", or "Agree" combined. (Additional response options not shown included "Strongly Disagree", "Disagree", or "Somewhat Disagree".)

| **Question: "To what extent do you agree or disagree with the following statements?"** | **"Strongly agree," "Somewhat agree" or "Agree" (%)** |
|---|---|
| I plan to pursue additional data analytics courses/experiences. | 85 |
| I want to continue doing research projects involving data analytics. | 74 |
| I want to do research in other areas of mathematics or science besides data analytics. | 68 |
| Experience using health related data for class projects and assignments has made me aware of the value of data analytics to the health care field. | 80 |
| I like using health related data for class projects/assignments. | 76 |
| I want to take more classes about data analytics. | 72 |
| I want to learn more about the theory behind data analytics techniques. | 70 |
| I want to further improve my computer skills. | 82 |
| I want to learn more ways to apply data analytics. | 79 |
| I want to have a career in data analytics. | 62 |
| Understanding data analytics will be of value to me no matter my career path. | 81 |
| I want to help solve more real-world problems using data analytics methods. | 77 |

- 62% of students indicated that they definitely wanted to pursue a career in data analytics.
- 81% understood DA would be of value no matter their career path potentially indicating that they are, at a minimum, "educated consumers" of data.
- 80% of students expressed that using health data as a basis for projects and assignments made them aware of the value of data in health care. (This is an appropriate outcome since in 2020 DAR focused exclusively on health informatics and COVID-19 research.)

### 5.3 Impact of PBL Framework for Undergraduate Math Research Experiences

We hypothesize that the success of the Data INCITE Lab and Pipeline, and especially DAR, is greatly enhanced by using the PBL framework and incorporating the design features described to provide an authentic experiential learning format with instructors coaching teams of students working on each project. This was also the conclusion of a prior study on the DATUM program from 2013-2016 that used the paid student internship model which subsequently evolved into the Data INCITE Pipeline described herein [20].

## 6 Concluding Remarks and Implications

We have found that a low-barrier pipeline consisting of an on-ramp DA course followed by DAR is a very effective method for incorporating DA learning and research into the undergraduate curriculum, providing an intriguing vehicle for: 1) attracting students to DA and related STEM studies; 2) better preparing and motivating students for further studies in mathematics and applying mathematics to any discipline; 3) teaching creative problem solving; 4) enabling meaningful undergraduate research; 5) teaching life-long data analysis skills; and 6) enhancing the competitiveness of graduates.

Our Data INCITE Pipeline includes an early course such as IDM that exposes the DA mathematical scaffold (i.e., the outlines and foundations of the techniques that let us express and find structure in high-dimensional data). IDM provides a common experience base that enables students to be highly productive at developing solutions to open problems in the DAR that follows. With IDM as a foundation, DAR is very effective at recruiting and developing deep analytical thinkers who can solve real-world DA problems. Our agile methodology also enabled us to quickly pivot to COVID-19 research. The students were able to apply their math, statistics, and coding skills to real "resume-building" work. We started COVID-19 research in April 2020, and devoted the full resources of the Data INCITE Lab to tackling COVID-19 related projects.

The method we have developed for running DAR as a CURE enables students to quickly immerse themselves in applied DA research and be immediately productive. Students' ability to complete a real-world project during DAR is enhanced by ensuring that a majority of students have basic skills in DA modeling and statistics, acquired through an introductory DA course with a PBL focus, and in some cases assembling a data preparation "advance team" to prepare the data for use during DAR. Paid or for-course-credit student interns work before and after DAR is in session to prepare datasets, to conduct preliminary research, and to finalize results, apps, and papers. This latter work is to ensure that valuable meeting time during the term is spent on creative research and not on the time-consuming process of finalizing papers and results for publication or apps for deployment. This partially accounts for the high productivity of the DAR students when the course is in session. We have already published three papers with focus on COVID-19 with more drafts in progress; we created multiple apps; and the work was presented by students and, at times, Dr. Bennett at major conferences.

Including DAR as part of the Data INCITE Pipeline reinforced the low-barrier access for DA learning and research, creating multiple entry options for all majors in addition to mathematics and computer science. We now allow students to enter the Data INCITE Pipeline with alternate R-based courses offered at RPI such as Biostatistics, Data Analytics, and Statistics serving as the on-ramp pre-requisites. We also encourage students in computer science machine learning courses, typically not R-based, to enroll in DAR as they usually adapt quickly to R. We note that any powerful DA language could be used as the base of the pipeline. For instance, we originally used MATLAB in IDM and then transitioned to R. Python, often the choice for introductory computer science courses and common in commercial DA environments, could be equally effective. We find that having students team with others from diverse backgrounds and skill sets actually boosts the productivity of DAR and is more closely representative of how teams work in the real world.

The success of the design and implementation of establishing DAR as a CURE was supported by a team that included experts in mathematics, DA, PBL pedagogy, and project management as was the case in DATUM [20]. We believe that any school that can assemble a team with this expertise should be able to replicate the success of our DAR and the Data INCITE Pipeline. RPI's strong traditions of undergraduate research and a talented, technically-oriented student body certainly also contribute to the success of DAR at our university. However, the basic ideas of the pipeline can be adjusted for other types of schools with different backgrounds and goals. For example, Siena College, a liberal arts school, has implemented their own version within their new data science program. They essentially divided IDM into a two-course sequence, *Introduction to Exploratory Data Analysis and Visualization* followed by *Introduction to Data Mathematics*. Siena College has also created their own CURE, *Data Science Team Project*.

The combination of an on-ramp course and DAR offers a model for developing and implementing curricular change for learning and research in mathematics using PBL. DAR proved to be a very compelling vehicle to engage early undergraduate mathematics students and enable them to be creative researchers solving real problems. Introducing high-dimensional mathematics in the context of DA earlier in the mathematics curriculum and providing early CUREs could be a more effective way to teach mathematics and attract students to STEM. Virtually all students experience the impact of DA in their lives on a daily basis (e.g., search engines, movie recommendations, shopping, and COVID-19 statistics). They experience high-dimensional data manipulation routinely. First-year undergraduates already have an intuitive sense of high-dimensional data and its analysis, yet it typically takes years for the standard mathematics curriculum to progressively develop the mathematical machinery necessary for high-dimensional modeling. The undergraduate mathematics curriculum could be much more engaging to students if it introduced and explored high-dimensional mathematical modeling and the profound impacts of DA on students' daily lives much earlier. High-dimensional mathematics including linear algebra and geometry and their application to the representation and manipulation of high-dimensional data can and must become a fundamental part of the early undergraduate experience. Students can develop the ability to extract information from data using mathematics, build critical skills that can be leveraged in future classes and later in life, and be recruited to further studies and careers in STEM.

This paper underscores how DAR design and infrastructure support creative research and self-directed learning for undergraduate students, creatively invoking their mathematical skills and engaging them in timely real-life projects with real-life clients. The success of DAR serves to re-affirm that as the future researchers and/or decision-makers, all students can benefit from at least a basic understanding of DA and that access to careers in DA can be increased through a low-barrier pipeline and CUREs that are inclusive by design.

**Acknowledgment.** *This work was supported by National Science Foundation Grant 1331023, United Health Foundation Grant 1990, and the Rensselaer Institute for Data Exploration and Applications (The Rensselaer IDEA).*

# References

[1] Ahrq rewards innovation in social determinants data visualization. `https://www.ahrq.gov/news/newsroom/press-releases/sdoh-challenge-rewards-innovation.html`. Accessed: 2021-06-21.

[2] github.com. `http://github.com`. Accessed: 2021-06-21.

[3] slack.com. `http://slack.com`. Accessed: 2021-06-21.

[4] Kristin P. Bennett. Artificial intelligence for public health. Keynote, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019.

[5] Kristin P. Bennett. Tackling COVID-19 using data analytics. `https://idea.rpi.edu/media/podcasts-and-videos/tackling-covid-19-using-data-analytics`, 2020. Institute for Data Exploration and Applications Video.

[6] Kristin P. Bennett, Lilian Ngweta, Karan Bhanot, and John S. Erickson. MortalityMinder: A web tool for visualizing and investigating social determinants of premature mortality in the united states. Systems Demonstration, AMIA 2020: American Medical Informatics Association Annual Symposium, Virtual Event, 2020.

[7] Kristin P. Bennett and Bruce R. Piper. EXTREEMS-QED: Data Analytics Throughout Undergraduate Mathematics (NSF Award 1331023). `https://bit.ly/3gEL8WF`, September 2013.

[8] Karan Bhanot, Syke Jacobson Jocelyn McConnon, L. Ngweta, and Kristin P. Bennett John S. Erickson. Investigating social determinants of premature mortality in the united states. Manuscript in revision, 2021.

[9] Winston Chang, Joe Cheng, J.J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2021. R package version 1.6.0.

[10] Chuoxi Chen, Mia Mayerhofer, Roma Paranjpe, Roma Paranjpe, Ruoyi Zhan, John S. Erickson, and Kristin P. Bennett. SafeCampus: Understanding population density using wifi data, 2021. Presentation, ACM New York Celebration of Women in Computing Conference.

[11] Shayom Debopadhaya, John S. Erickson, and Kristin P. Bennett. Temporal analysis of social determinants associated with COVID-19 mortality. *ACM BCB '21: 12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 41:1–41:10, 2021.

[12] Shayom Debopadhaya, Ariella D. Sprague, Hongxi Mou, Tiburon L Benavides, Sarah M. Ahn, Cole A. Reschke, John S. Erickson, and Kristin P. Bennett. Social determinants associated with COVID-19 mortality in the United States. *medRxiv*, 2020. Poster Presentation at AMIA 2020 by S. Debopadhaya. Also invited presentation at Virtual International COVID Diabetes Summit 2020 by Dr. Bennett.

[13] Weinan E. The dawning of a new era in applied mathematics. *Notices of the American Mathematical Society*, 68(4):565–571, 2021.

[14] Sophie Hannigan, Christina van Hal, Kristin P. Bennett, and John S. Erickson. MortalityMinder: Diving into the causes and trends of mortality, 2020. Presentation, ACM New York Celebration of Women in Computing Conference.

[15] Hannah De los Santos. ECHO (Extended Circadian Harmonic Oscillators) github. `https://github.com/delosh653/ECHO`, 2020. See also: https://doi.org/10.1093/bioinformatics/btz617.

[16] Hannah De los Santos, Shayom Debopadhaya, Kaelyn M. Edwards, Alexandra M. David, Uyen H. Dao, Kristin P. Bennett, and Jennifer M. Hurley. The PAICE suite identifies circadian oscillations in non-coding macrophage RNAs, 2021. Manuscript in preparation for submission.

[17] Rufeng Ma and Rachael C. White. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 twitter discourse, 2021. Presentation, ACM New York Celebration of Women in Computing Conference.

[18] National Academies of Sciences, Engineering, and Medicine and others. *Data science for undergraduates: Opportunities and options.* National Academies Press, 2018.

[19] Abraham Sanders, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S. Erickson, and Kristin P. Bennett. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse. AMIA, 2020.

[20] J.C. Seddon. *Project-based learning: A case study of early data analytics learning in undergraduate mathematics.* PhD thesis, University of Toronto, Toronto, Canada, 2021.

[21] Scott W. Slough and John O. Milam. Theoretical framework for the design of STEM project-based learning. In *STEM Project-Based Learning*, pages 15–27. Brill Sense, 2013.

[22] The Rensselaer IDEA Data INCITE Team. Mortalityminder github. `https://github.com/TheRensselaerIDEA/MortalityMinder`, 2019.

[23] The Rensselaer IDEA Data INCITE Team. MortalityMinder web app. `https://inciteprojects.idea.rpi.edu/apps/mortalityminder/`, 2019.

[24] The Rensselaer IDEA Data INCITE Team. COVID back-to-school web app. `https://inciteprojects.idea.rpi.edu/apps/backtoschool/`, 2020.

[25] The Rensselaer IDEA Data INCITE Team. COVID masks nlp github. `https://github.com/TheRensselaerIDEA/COVID-masks-nlp`, 2020.

[26] The Rensselaer IDEA Data INCITE Team. COVID Twitter github. `https://github.com/TheRensselaerIDEA/COVID-Twitter`, 2020.

[27] The Rensselaer IDEA Data INCITE Team. COVID WarRoom github. `https://github.rpi.edu/DataINCITE/IDEA-COVID-WarRoom`, 2020. Includes code for COVID WarRoom and COVID apps.

[28] The Rensselaer IDEA Data INCITE Team. COVID warroom web app. `https://inciteprojects.idea.rpi.edu/apps/warroom/`, 2020.

[29] The Rensselaer IDEA Data INCITE Team. COVIDMINDER github. `https://github.com/TheRensselaerIDEA/COVIDMINDER`, 2020.

[30] The Rensselaer IDEA Data INCITE Team. COVIDMINDER web app. `https://inciteprojects.idea.rpi.edu/apps/covidminder/`, 2020.

[31] The Rensselaer IDEA Data INCITE Team. Rpi SafeCampus github. `https://github.rpi.edu/DataINCITE/IDEA-COVID-SafeCampus`, 2020.

[32] The Rensselaer IDEA Data INCITE Team. RPI StudySafe github. `https://github.rpi.edu/DataINCITE/IDEA-COVID-StudySafe`, 2020.

[33] The Rensselaer IDEA Data INCITE Team. SafeCampus web app. `https://inciteprojects.idea.rpi.edu/apps/safecampus/`, 2020.

[34] The Rensselaer IDEA Data INCITE Team. Social Determinants Associated with COVID-19 Mortality github. `https://github.com/TheRensselaerIDEA/COVIDMINDER/tree/master/social-determinants-paper`, 2020. See also: https://www.medrxiv.org/node/95535.external-links.html.

[35] The Rensselaer IDEA Data INCITE Team. StudySafe web app. `https://inciteprojects.idea.rpi.edu/studysafe/app/studysafe`, 2020.

[36] Xiao Wu, Rachel C. Nethery, Benjamin M Sabath, Danielle Braun, and Francesca Dominici. Exposure to air pollution and COVID-19 mortality in the United States. *medRxiv*, 2020.

[37] Jennifer Y. Zhang, Trisha Shang, David Ahn, Kong Chen, Gerard Coté, Juan Espinoza, Carlos E. Mendez, Elias K. Spanakis, Bithika Thompson, Amisha Wallia, et al. How to best protect people with diabetes from the impact of sars-cov-2: Report of the international COVID-19 and diabetes summit. *Journal of Diabetes Science and Technology*, 15(2):478–514, 2021.