A CS1 Open Data Analysis Project with Embedded Ethics

Shira Wein
sw1158@georgetown.edu
Georgetown University
Washington, D.C., USA

Alicia Patterson
alicia.patterson@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Shannon Brick
sb2124@georgetown.edu
Georgetown University
Washington, D.C., USA

Sydney Luken
Sydney.Luken@georgetown.edu
Georgetown University
Washington, D.C., USA

Course CS1
Programming Language Python
Knowledge Unit Programming Concepts
CS Topics Functions, Data Types, File Handling, Data Visualization, Data Analysis, If Statements, Loops
Resource Type Project

SYNOPSIS

This final project combines key CS1 programming concepts with ethical analysis. It helps students gain experience with lists, dictionaries, for/while loops, conditional statements, file handling, and functions in Python. Through a data analysis and visualization task, the students put to action their prior knowledge of the aforementioned programming concepts, embedded with an ethics-led discussion of *open source data*. Open source data (or "open data") is data that is available and accessible to anyone, including for reuse of the data [8]. Students will learn how to think critically about the ethical dimensions of their selected open source data (and future open source data), and provide an analysis of the data within its contemporary cultural context.

KEYWORDS

open source, data ethics, introductory, non-majors

ACM Reference Format:
Shira Wein, Alicia Patterson, Shannon Brick, and Sydney Luken. November 2023. A CS1 Open Data Analysis Project with Embedded Ethics. In *ACM EngageCSEdu.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3631987

ACM EngageCSEdu, November 2023.

ACM ISBN 979-8-4007-0468-0/23/11.
https://doi.org/10.1145/3631987

1 INTRODUCTION

Embedding ethics across the computer science degree increases the likelihood that students who take only one computer science gain some appreciation for the ethical dimensions of computer science. These are, for instance, undergraduate non-majors or high school students who elect to take a computer science course to fulfill a requirement or to develop a supplemental skill on the job search. Though these students have limited access to (and knowledge of) the tenets of responsible computing, they are liable to use and develop code in the future, and thus would still very much benefit from computational ethics training.

Our module for the final project of an introductory undergraduate class integrates ethical analysis and assessment of open source data with Python programming and data analysis. Our ethics embed asks students to consider: the perspectives and assumptions present in datasets, the benefits and harms associated with open source data, and ethical issues related to re-using data (including informed consent, autonomy, and data privacy).

2 ENGAGEMENT HIGHLIGHTS

The purpose of the final project is to have students implement a larger-scale solution to a problem which integrates their knowledge from the other units in the course. The project requires the use of file handling, conditionals, loops, string manipulation, dictionaries, lists, and functions. This should also serve as a fun way for students to integrate their own interests (by way of the specific file and data analyzed) with their new Python skills. Each group chooses one open source dataset from the provided options. Options of the data in our implementation included: descriptions of satellites orbiting Earth, descriptions of US mass shootings, drug use by age, college majors by gender and employment rate, congress resignations by year and party, yearly greenhouse gas emissions by country, and World Corruption Index data, among others. The variety of data options **incorporates**

student choice and promotes **interdisciplinary connections to CS**, between the students' major fields and their programming skills.

Upon selecting their datasets (by voting for their preferences in an online poll), the groups embark on a series of pre-class, in-class, and post-class activities as part of the ethics module. The pre-class work had students research basic facts about the origins of their datasets. The in-class session was led by ethicists and included both a brief lecture component on open source data as well as discussions of the ethical considerations relevant to the specific open source datasets used in the students' final projects. Each aspect of this project is group-oriented in order to thread **well-structured collaborative learning**, though each of the three required analyses can be done individually (one analysis per person) or collectively (with all students contributing to multiple analyses).

Using their selected datasets, the students (1) take in a file, (2) clean and process the data (handling any missing data), and then (3) produce three analyses of the data, by funneling the relevant information into the necessary data types.

The end result of this module is that, in addition to applying various technical concepts in the context of a larger-scale project, students gain a greater appreciation of the ethical dimensions of open source data, and are more prepared to engage with such data responsibly in the future.

3 ETHICS LEARNING GOALS

The overarching learning goal of the ethics component was to give students tools to think critically about their datasets. More specifically, we aimed to enable students to:

- Grasp the normative role of consent, and what it means to meaningfully consent to the collection and use of one's data.
- Appreciate the difficulty of meaningfully consenting to the re-use of one's data, and how this difficulty can challenge the acceptability of making data open source in the first place.
- Begin to appreciate the extent to which the creation of data requires that humans make decisions about how to define, categorize, and standardize the things they are studying [2]; to be prepared to ask about how these decisions might impact the use of data in ways that are more or less harmful/beneficial for different groups; to think critically about what data maybe missing or excluded as a result of how the data has been categorized or standardized; to understand that similar decisions related to data analysis must often be made by users of data sets, in particular when they engage in the process of cleaning the data.
- Begin to develop their own ethical agency as users and consumers of data, by equipping them with specific questions they can ask of any data set they encounter in the future, in order to make a responsible decision about how best to use it.
- Think creatively about how the norms of open source data could change to better facilitate responsible use of open source data.

4 MODULE IMPLEMENTATION

The role of the instructor is to curate the in-class lecture material and select which open source datasets the students can use. The instructor will need to (1) assign the project in the course and the pre-class work, (2) either lead the in-class ethics lecture themselves or collaborate with an ethicist on campus to lead that session, [1] and (3) assess the post-class work and final project submission.

4.1 Project Assignment

The data analysis project was assigned as the final project, a month before the final deadline (though the data science project could easily be adapted to a different scope or difficulty).

For the final projects, the students work in groups of 3 (approximately) to create a codebase that takes in a CSV of open source data and analyzes it. Specifically, the students (1) clean and process the data (handling any missing data or outliers), and (2) every member of the group creates a function to analyze the data. During an in-class lecture on data analysis and data visualization, which took place early on during the time period students were working on the project, the instructor discussed techniques and considerations for preprocessing and cleaning up noisy data. Specifically, duplicate data points, missing values, and outliers were discussed, with regard to how they can be detected and whether you may want to handle them during preprocessing. For example, regarding outliers in the case of analyzing salary values by major, the students should decide whether an outlier causes an inaccurate picture based on the statistic they are using (e.g. if there is one value with a salary of $0), but must be careful not to remove an excessive number of outliers which provide important information about the range of the data (e.g. if the data point indicates unemployment after graduation). This can require more investigation on the part of the student in order to ascertain whether the outlier should be included or not in the average statistics, but additional

[1] For helpful tips on classroom collaborations with other discipline, see "Working Across Disciplines" in Mozilla's *Teaching Responsible Computing Playbook* at https://foundation.mozilla.org/en/what-we-fund/awards/teaching-responsible-computing-playbook/topics/working-across-disciplines/.

information may not be available. The students then needed to discuss their choices for preprocessing and cleaning the data during their final presentation and in their submission.

Some examples of potential functions might be creating a list of all the values in a column, for which you could then take the mean or max value if it is integers (numerical data), or find the most common value if it contains strings (categorical data). Another option is an interactive function which asks a user for input about a specific value, e.g. how many times a certain value occurs in the data. Finally, a data visualization using `matplotlib` could represent the data, such as a line graph which shows results over time.

Because this was the first time students would be asked to engage with data from an ethical perspective, it was important to select data sets whose susceptibility to ethical analysis and assessment was relatively easy to spot. For example, a dataset on dry beans is harder for undergraduates at an introductory level to notice ethical issues. Datasets that relate to social, political issues broadly construed will be the least challenging. The data was primarily retrieved from GitHub[2] and Kaggle[3].

The project description includes details of the post-class requirements, which are to be included in the final project submission.

Though we introduced these foundational ethics concepts of open data (e.g. data sharing consent, data provenance, bias in data collection) within the context of the final project, it would also be possible to introduce this issue separately, earlier in the course. Then, these topics could be expanded throughout the course and in the final project. Students could complete readings on these topics to prepare them for these discussions ahead of time, such as selections from two books, both of which are designed for computer science students: Baase [1] which introduces computing ethics broadly and considers a handful ethical and legal issues (e.g. privacy, intellectual property, and risk assessment) or Burton et al. [3], which couples formal discussion of ethical frameworks with science-fiction case studies presenting ethical dilemmas. Additional resources which students could read throughout/earlier in the course include pages 130-137 in D'ignazio and Klein [4], Radin [13], Peng et al. [11], and Metcalf and Crawford [7].

4.2 Pre-class Work

Before the in-class ethics session, each student group completed the first "Data Sheets for Data Strangers" worksheet. Inspired by Gebru et al. [5] as well as Catherine D'Ignazio and Lauren Klein's notion of "strangers in the dataset," [4], this worksheet prompts the students to identify:

- **WHO** Who is this data about? Who collected this data? Who is this dataset for / who is it intended to benefit or serve?
- **WHAT** What is the data about? Which attributes or data are ambiguous? Are there any attributes for which an explicit definition is provided? [4]
- **WHEN** What timespan is represented in the data? When was the data collected?
- **WHERE** Where is the data from? (i.e. does it pertain to residents of a specific area or geographic location?)
- **WHY** Why was this data collected? What is its intended use?
- **HOW** How was the data collected?

Students were thus prompted to think critically about what they do know, and what they cannot know, about the history of their dataset. The purpose of this exercise was to enable students to begin to grasp the extent to which we are often ignorant of potentially important information about open source data, and what we need to know about data to responsibly re-use it.

4.3 In-class Session & Work

In class, a session led by ethicists included a lecture on the ethics of open source data. Then, the session leaders aided students in thinking about the ethics of the open source data they are using for their final project as students completed a worksheet in their project groups.

The lecture focused on two sets of issues for open source data: 1) the importance of participant consent to the collection, use and reuse of their data, and; 2) the ways in which the construction of a data set necessarily renders certain assumptions and perspectives visible or invisible through the choices of how to categorize and standardize a subject into data fields [9, 10], in ways that can carry significant consequences for different populations of people. The discussion of both sets of issues foregrounded the need for computer scientists to know more about **data provenance** in order to make responsible decisions about what appropriate use looks like.

In response, this enables students to begin thinking about the ethical dimensions of open source, and be prepared with specific questions in mind to delve into a range of ethical issues for their selected dataset. Discussed issues including consent (such as in the case of collecting health information [15] and video surveillance [14]), invisibility of certain populations in data collection [12], the assumptions baked

[2]https://github.com/
[3]https://www.kaggle.com/

[4]We would recommend additionally considering the inclusion of questions which target *how* students should use the dataset, as opposed to describing the dataset. For example, this could take the form of: "WHAT What might need to be changed, in the process of cleaning the data and preparing it for analysis? HOW How might these changes impact what the data represents?"

into dataset creation [6], and how to create processes for the ethical use of open data [5].

With the new knowledge gleaned from the lecture, the students then filled out and discussed the second "Data Sheets for Data Strangers" worksheet. The students considered the topics from the lecture with respect to their own data for the final project. This worksheet contained four topics of questions, and students were encouraged to focus on the questions that struck them as most relevant to their own data set. Here, we list the four topics and some examples of questions included under those four topics:

- **CONSENT** e.g. Did the subjects consent to the data being collected initially?
- **DIGGING INTO DEFINITIONS** e.g. Will some of these definitions include/exclude different people?
- **CATEGORY CONSTRUCTION** e.g. How might the dataset reflect the assumptions, motivations, and interests of its creator(s)?
- **BENEFITS AND HARMS** e.g. Is this data about people who might be especially vulnerable? Can the data be put to use in a way that might harm the data subjects?

4.4 Post-class work

After the in-class session, the students were required to incorporate into the (approx. 2 page) final project document an ethics statement about their data.

The students were tasked with writing 1-2 paragraphs about an ethical issue/concern that is most relevant to their dataset. To do this, the students needed to identify an ethical issue and explain how this issue is specifically relevant to their dataset, plus what considerations are necessary for the use of this data.

The final project submission with code, documentation, and slides, were submitted online, and all students presented their work in class towards the end of the semester.

5 RECOMMENDATIONS

Our implementation benefited from collaborative development between the ethicists and the instructor of the computer science course. Over the course of some months, the computer science instructor met with ethicists and a designer to create an activity-based class on the ethics of open source data, accompanied by assignments, all of which prepared students to address an ethical issue connected to data that was analyzed in the course's final project. Having a collaborator on-campus with a background in responsible computing and/or ethics generally lead the in-class session will reduce the workload for the instructor and ensure a high-quality lecture with up-to-date ethical considerations, though having a collaborator is not required. The materials provided here are the ones we developed, so small changes to the lecture and dataset options could certainly be done by the instructor alone.

An instructor who implements this project will need to curate a list of relevant datasets for the students to choose from. In order to select open data sets that will be effective for this project, it is important to consider: (1) the appropriateness of the content of the dataset for the age group, (2) whether the dataset is well-documented and well-organized from a technical standpoint, and (3) the ethical richness of the content. It is important that the instructor is sensitive to the possibility that some data subjects might be particularly sensitive and/or difficult for students, and students should therefore be able to choose for themselves which data sets they will deal with (i.e., a student who might have grown up in closer proximity to gun violence, should not be forced to work on a project about gun violence.)

Asking the students about their majors or values will enable the instructor to include options that could be of interest to the students in the class. The open source data choices should also be selected such that they could encourage discussion of ethical considerations within the class period. For example, data about congress resignations and census-collected data on college majors by gender and employment rate sparked rich ethical analysis in discussion, while data on US births and artificial data on employee attrition were more difficult for the students to analyze. Data that is well-documented includes text describing how the data was collected and what the columns/rows in the spreadsheet correspond to.

Two good sources of well-documented and well-organized open data are through FiveThirtyEight's GitHub, which contains all of the data used for data analysis and visualization on FiveThirtyEight's website [5] or Kaggle's datasets [6]. One of the benefits of using Kaggle is that it scores the "usability" of the dataset, which indicates how credible and platform-compatible the data is. Also, some datasets in Kaggle are web scraped (e.g. the OkCupid dataset at https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles), which raises further questions around consent and the ethics of data re-use. The FiveThirtyEight data has usually been used in an existing analysis or visualization, which can be either to the benefit or detriment to students. There will likely be code or analyses of the data currently existing, which can serve as an interesting comparison or starting point for the student's analysis, but it will be important for the instructor to ensure that it does not interfere with the integrity of the student's submission. An additional resource is the "Data is Plural" archive, which contains datasets which have already been

[5] https://github.com/fivethirtyeight
[6] https://www.kaggle.com/datasets

cleaned (at https://data.world/jsvine/data-is-plural-archive). An additional resource is the "Data is Plural" archive, which contains datasets which have already been cleaned (at https://data.world/jsvine/data-is-plural-archive).

Before assigning the datasets to students, it is important for the instructor to first download the data and verify that it is well-formed and free to access. Then, the instructor can send the students the data themselves as opposed to requiring the student to download it online. Still, in order to complete the in-class work, the students will need to explore the website of origin in order to identify the characteristics of the dataset. We have included in the materials the data options we presented to students as a starting point; this should be modified to match the interests, computational abilities, and age of the students.

This project would be appropriate for either high school or undergraduate students in their first computer science class (i.e. CS1), as the ethics concepts do not require prior experience and are sufficiently straightforward.

Having this project be the final project allowed students to draw from many relevant technical concepts in their final implementation, including data types, file handling, functions, and data visualization/analysis. By embedding the ethical analysis and assessment into a larger body of work––rather than having, say, a one-off lecture on data ethics–students were encouraged to recognize the kinds of questions that should be asked in the course of any ethically responsible data science project.

The content and structure of the lecture can also be shifted to highlight ethical issues most relevant to the project datasets. In our implementation (in a college setting), the students really gravitated towards the topic of *consent*; this reflects in part the fact that the first topic covered during the in-class session was that of consent and how that relates to dataset development. In a future iteration, we intend to present the questions in a non-linear format in order to de-emphasize consent as the primary ethical concern.

The scope of the programming piece can also be scaled depending on how much time you are dedicating to the final project (we allotted a month between the assignment of the project and its due date).

6 STUDENT FEEDBACK

In our implementation of this module, students remarked on the fact that prior to the module they had never considered that there were ethics or responsibility embedded into computer science. Other students came away from the final project stating that evaluating the ethical considerations for this project impacted their data consumption habits in their personal lives and that they have more tools for thinking about how to avoid harm in any future programming work. While this was a nice outcome from our implementation of this module, it would also be possible to prompt this thinking in students by specifically asking them how the ethical components of the module might shape their future interactions with open source datasets, and data more generally.

Relating to the ethical concerns present in their dataset, many students noted inherent bias in the data (e.g. gender being coded as binary, not including any non-binary), the sensitivity of the data, and potential issues with the collection of the data itself (e.g. the students were not sure whether respondents were told that they had a right *not* to respond to the questionnaires which generated the data for their dataset).

7 MATERIALS

- "Final Project" project assignment and grading rubric
- One-page "Data Sheets for Data Strangers" pre-class worksheet
- Two-page "Data Sheets for Data Strangers" in-class worksheet
- "The Ethics of *Open* Data" sample lecture slides
- Data Options sample sheet

REFERENCES

[1] Sara Baase. 2012. *A gift of fire.* Pearson Education Limited.
[2] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences.* MIT press.
[3] Emanuelle Burton, Judy Goldsmith, Nicholas Mattei, Cory Siler, and Sara-Jo Swiatek. 2023. *Computing and Technology Ethics: Engaging through Science Fiction.* MIT Press.
[4] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.
[5] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010
[6] Manissa M Maharawal and Erin McElroy. 2018. The anti-eviction mapping project: Counter mapping and oral history toward bay area housing justice. *Annals of the American Association of Geographers* 108, 2 (2018), 380–389.
[7] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211.
[8] Jennifer C Molloy. 2011. The open knowledge foundation: open data means better science. *PLoS biology* 9, 12 (2011), e1001195.
[9] Mimi Onuoha. 2016. The Library of Missing Datasets. Mixed-media installation. https://mimionuoha.com/the-library-of-missing-datasets
[10] Mimi Onuoha. 2016. The point of collection. *Data & Society: Points* (2016).
[11] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922* (2021).
[12] Tonia Poteat, Danielle German, and Colin Flynn. 2016. The conflation of gender and sex: Gaps and opportunities in HIV data among transgender women and MSM. *Global public health* 11, 7-8 (2016), 835–848.

[13] Joanna Radin. 2017. "Digital Natives": how medical and indigenous histories matter for big data. *Osiris* 32, 1 (2017), 43–64.

[14] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*. Springer, 17–35.

[15] Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association, 261.