CS 598 Foundations of Data Curation Project

Usman Asghar (uasgh2@illinois.edu)
Hasham Ul Haq (huhaq2@illinois.edu)
Jiaqing Mo (jiaqing7@illinois.edu)

Overview

The overall goal of this project is to curate a dataset that explores the relationship between student well-being and academic performance in college. Student well-being has become an important focus in higher education, as stress, sleep, social life, and mental health can all influence how students perform academically. This dataset will provide a structured foundation for studying how different aspects of well-being connect to academic outcomes.

The use case is to enable researchers, educators, and institutions to better understand the role of well-being in student success. With this curated dataset questions such as How do sleep patterns relate to GPA? Does physical activity or social engagement impact academic performance? Which well-being factors have the strongest correlation with outcomes? can be addressed. Ultimately the project will contribute to building a clearer picture of how student lifestyle and well-being shape educational performance.

1. Data Lifecycle

Applying the USGS Science Data Lifecycle Model (SDLM)

The USGS Science Data Lifecycle Model provides a linear but iterative view of scientific data management with six major elements—Plan, Acquire, Process, Analyze, Preserve, and Publish/Share—and several cross-cutting activities like metadata documentation, quality management, and security. Each stage maps directly onto aspects of the StudentLife project.

1. Plan

Planning in StudentLife involved defining research goals (understanding student well-being through digital sensing), determining the scope of data collection, obtaining ethical clearance, and designing consent procedures. This aligns with the SDLM's emphasis on considering activities "from inception through to archiving." Key planning concerns included participant privacy, data anonymization protocols, sensor sampling frequency, and storage infrastructure. The planning phase also involved cross-disciplinary

coordination among computer scientists, psychologists, and data engineers to ensure that technical capabilities supported the research objectives.

2. Acquire

The acquire phase involved collecting multimodal data via smartphones and institutional systems. Sensors continuously captured contextual signals, while Ecological Momentary Assessments (EMAs) provided subjective inputs. This stage required balancing data fidelity with participant burden and power consumption. The SDLM also stresses the importance of metadata at this point—StudentLife documented timestamps, sensor types, and sampling configurations, enabling reproducibility and data quality assessment later.

3. Process

Processing involved cleaning raw data, handling missing values, synchronizing streams, and removing personally identifiable information (PII). For instance, GPS coordinates were generalized to protect location privacy, and audio snippets were processed to extract features rather than store raw recordings. These align with the SDLM's processing focus on preparing data for analysis while ensuring integrity and quality control.

4. Analyze

The analysis phase centered on linking behavioral features (sleep, mobility, sociability) to academic outcomes and mental health measures. Researchers used machine learning and statistical modeling to uncover patterns of stress and performance. According to the SDLM, this stage translates processed data into scientific understanding, requiring close attention to provenance and reproducibility.

5. Preserve

Once analysis was complete, the StudentLife dataset required long-term preservation strategies. Raw and derived data were stored in standardized formats, accompanied by documentation describing collection procedures and feature extraction methods. Preservation ensured that future researchers could reuse the data without ambiguity, consistent with SDLM guidance.

6. Publish/Share

The dataset's release under controlled access reflects the final publication/sharing phase. The USGS model stresses dissemination with proper citation and licensing. StudentLife followed similar principles: sharing

anonymized data to advance research while maintaining participant confidentiality through restricted access requests and ethical review.

Cross-Cutting Activities

Throughout all stages, cross-cutting elements such as metadata documentation, quality management, and security were integral. StudentLife maintained structured metadata for each sensor and survey, implemented scripts for validation, and enforced strict security measures to prevent data breaches. These actions illustrate how cross-cutting processes ensure data reliability across lifecycle stages.

2. Ethical, legal, and policy constraints

1. Ethical Constraints

Respect for Persons: Informed Consent and Autonomy

The Belmont Report defines "respect for persons" as the obligation to treat individuals as autonomous agents and to protect those with diminished autonomy (National Commission for the Protection of Human Subjects, 1979). The Menlo Report extends this principle to information and communications technology (ICT) research, emphasizing the importance of informed consent and clear communication (Department of Homeland Security, 2012).

In the StudentLife study, all participants were fully informed about what data were collected—ranging from GPS and audio features to mood surveys—and how those data would be used and stored. Participants voluntarily consented through an institutional review board (IRB)–approved process, with the right to withdraw at any time. This transparent consent process ensured that participation was both voluntary and informed, directly satisfying the Belmont and Menlo principles of respect for persons.

Beneficence: Minimizing Harm and Maximizing Benefit

The principle of beneficence, also derived from the Belmont Report, requires that researchers "do no harm" and minimize potential risks while maximizing potential benefits (Resnik, 2020). The StudentLife dataset involved continuous monitoring of participants' movements and behaviors, creating risks of psychological discomfort or exposure if re-identification occurred. To mitigate these risks, researchers employed multiple data protection strategies—including de-identification, encryption, and aggregation of data to reduce granularity.

For example, raw GPS coordinates were replaced with generalized regions, and audio data were transformed into non-speech features rather than stored recordings. These measures minimized the potential for harm while allowing the data to support valuable research on student well-being. The result reflects the Menlo Report's emphasis on balancing individual risks against collective societal benefits (Department of Homeland Security, 2012).

Justice: Fairness and Non-Discrimination

The principle of justice calls for fairness in both the selection of research subjects and the distribution of benefits (National Commission, 1979). In StudentLife, all participants were drawn from a single university population, preventing bias in data inclusion. The dataset's availability under controlled-access conditions ensures that the research benefits extend broadly—to psychologists, computer scientists, and behavioral scientists—without compromising participant welfare.

This balance aligns with the Menlo Report's interpretation of justice, which calls for equitable treatment and fair distribution of research burdens and benefits (Department of Homeland Security, 2012).

Respect for Law and Public Interest

The Menlo Report introduces Respect for Law and Public Interest as a fourth principle, requiring compliance with legal standards and accountability to the public (Department of Homeland Security, 2012). Dartmouth researchers ensured compliance with all relevant laws governing human subjects research, including federal policies like the Common Rule (45 CFR 46). They also documented their methods, consent forms, and data processing steps transparently, allowing public scrutiny and replicability. This accountability upholds the Menlo principle of transparency and the Belmont principle of integrity in research.

2. Legal Constraints

Human Subjects Research and the Common Rule

Under the Common Rule, human subjects research includes any systematic investigation that collects identifiable private information (U.S. Department of Health and Human Services, 2018). The StudentLife project qualified under this definition due to its collection of behavioral and mental health data.

Accordingly, the study underwent IRB review to ensure that data collection methods, informed consent procedures, and data storage complied with ethical and legal requirements. By securing IRB approval and implementing anonymization before data release, the project demonstrated compliance with the federal policy for protecting human subjects.

Privacy Laws: HIPAA, FERPA, and GDPR

The StudentLife dataset touches on multiple privacy regulations, even though it does not explicitly fall under medical or European jurisdiction.

The Health Insurance Portability and Accountability Act (HIPAA) establishes privacy protections for identifiable health information (U.S. Department of Health and Human Services, 2003). Although StudentLife was not a medical study, it collected mental health–related survey responses and sleep data that could be interpreted as health information. Researchers followed HIPAA's de-identification guidelines by removing or transforming all 18 categories of protected health identifiers.

The Family Educational Rights and Privacy Act (FERPA) governs educational records in the United States (U.S. Department of Education, 1974). Since StudentLife also captured academic outcomes (e.g., GPA, attendance), FERPA's restrictions on disclosing educational identifiers were respected through anonymization and secure data handling.

Finally, the General Data Protection Regulation (GDPR)—though an EU law—has influenced global privacy standards. GDPR's principles of data minimization, purpose limitation, and accountability (GDPR, 2018) are reflected in StudentLife's practices: researchers collected only necessary data, stored them securely, and released them under restricted access with documented consent.

De-Identification and Risk Reduction

As noted in the CS598 ethics module, protecting privacy often requires a balance between data utility and confidentiality (Garfinkel et al., 2023). The StudentLife dataset adopted multiple de-identification techniques consistent with those described in the De-Identification Methods section of the course:

- Suppression: Removal of direct identifiers such as names and phone numbers.
- Generalization: Aggregating GPS data to neighborhood-level regions.
- Perturbation: Random shifts in time or location data.
- Pseudonymization: Replacement of user IDs with random identifiers.

- k-Anonymity: Ensuring that individual records were indistinguishable from others within groups.

These strategies collectively minimized re-identification risks while maintaining analytical integrity.

3. Policy Constraints

Institutional and Federal Data Policies

Federal research agencies increasingly require data management plans that ensure both openness and ethical handling. The 2022 OSTP "Nelson Memo" and 2023 NIH Data Management and Sharing Policy mandate that federally funded research data be shared publicly when possible, while protecting confidentiality (Office of Science and Technology Policy, 2022; NIH, 2023).

The StudentLife dataset aligns with these expectations: it provides open access to anonymized data and metadata through a controlled repository, supporting the FAIR (Findable, Accessible, Interoperable, Reusable) principles. However, to prevent misuse, researchers must apply for access, agree to data-use terms, and acknowledge the original study—reflecting both ethical openness and legal restraint.

Licensing and Intellectual Property

Although raw data are generally not subject to copyright protection, data compilations and arrangements can be licensed (Carroll, 2009). StudentLife's creators released the dataset under a research-use license, restricting commercial exploitation and requiring attribution. This follows the Respect for Law and Public Interest principle, ensuring that the data remain a public good used for legitimate scientific inquiry.

Transparency and Accountability

Transparency and accountability are continuous responsibilities throughout the data lifecycle (Plale & Kouper, 2017). StudentLife's documentation includes descriptions of consent processes, data types, preprocessing steps, and ethical safeguards. These efforts mirror the course's emphasis on maintaining trust, integrity, and reproducibility across all data management stages (Resnik, 2020).

Integration with Data Lifecycle and Curation Ethics

The Research and Data Ethics framework emphasizes that ethical and legal responsibilities span every stage of the data lifecycle—from planning and acquisition to preservation and sharing. For StudentLife, this meant:

- Planning: Conducting ethical risk assessments and obtaining IRB approval.

- Acquisition: Ensuring participant consent and data minimization.
- Processing: Applying anonymization and encryption.
- Preservation: Storing data securely with access control.
- Publishing: Providing transparent documentation and controlled access for reuse.

Each stage demonstrates alignment with both ethical frameworks (Belmont, Menlo) and data governance models (USGS, DCC, SEAD) discussed in CS598.

Licensing and usage policy:
This dataset does require citation - we will cite the relevant paper in our project.
To keep it simple and avoid any further licensing challenges, we will not publish the data as part of our submission. Users can download the data themselves. We will be publishing the details of the dataset, scripts, and metadata - without the actual copy of the dataset.

3. Data models and abstractions

Each record is anchored by a consistent participant identifier (uid) that links multiple data sources, such as surveys, sensor logs, and academic records. Variables are explicitly typed - continuous (e.g., sleep hours, GPA), categorical (e.g., course name), or ordinal (e.g., stress levels)—to ensure consistency and interpretability during analysis. This relational structure allows us to map well-being indicators to academic outcomes through shared identifiers and time periods, supporting efficient joins, normalization, and correlation analyses. The schema was designed to promote clarity, reduce redundancy, and enable scalable integration as additional modalities (e.g., lifestyle or phone activity) are incorporated.

Schema Description

The schema follows a relational design centered around the participant–term relationship. Each participant (uid) may have multiple associated well-being and academic records per academic term

(term_id). Courses and enrollments extend this model to represent per-course performance. Relationships are defined using primary and foreign keys to maintain data integrity and facilitate reproducible joins across tables.

- Participants → stores demographic and identifier information.
- Terms → represents academic terms (e.g., 2013 Spring → 13S).
- Wellbeing Metrics → captures term-level indicators such as average sleep hours and stress levels.
- Academic Metrics → stores term and cumulative GPA along with attendance rates.
- Courses and Enrollments → record per-course grades linked to participants and terms.

In total, there are 30 unique students whose GPA information is available, and following is the distribution of that table:

	gpa all	gpa 13s	cs 65
count	30.000000	30.000000	30.000000
mean	3.421533	3.330556	3.622222
std	0.397754	0.798284	0.796224
min	2.400000	1.000000	0.000000
25%	3.257000	3.333333	3.666667
50%	3.490500	3.527778	4.000000
75%	3.698500	3.861111	4.000000
max	3.947000	4.000000	4.000000

This project uses a relational data model to organize well-being and academic data into linked tables. Each entity—Participant, Term, Well-Being Metrics, Academic Metrics, Course, and Enrollment—is represented as a table connected through shared identifiers (uid, term_id, course_id). This structure supports consistency, normalization, and efficient analysis across time and individuals.

Key abstractions include:

- Entity abstraction: defines participants, courses, and terms as distinct conceptual units.
- Identifier abstraction: uses uid and term_id to maintain linkage across data sources.
- Temporal abstraction: aggregates time-based measures (e.g., sleep hours, stress) by academic term.
- Attribute typing: continuous (GPA, sleep hours), ordinal (stress level), and categorical (major, course).

These abstractions enable integration of multi-modal data while preserving interpretability. The relational schema supports one-to-many and many-to-many relationships, allowing each participant to have multiple well-being and academic records per term, and multiple course enrollments.

References

Carroll, M. W. (2009). Licensing and contracting issues in databases. In Encyclopedia of Database Systems. Springer. https://doi.org/10.1007/978-0-387-39940-9_1518

Department of Homeland Security. (2012). The Menlo Report: Ethical principles guiding information and communication technology research. https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf

Garfinkel, S., Near, J., Dajani, A., Singer, P., & Guttman, B. (2023). De-identifying government datasets: Techniques and governance. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-188

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. U.S. Department of Health and Human Services.

National Institutes of Health. (2023). Final NIH policy for data management and sharing. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Office of Science and Technology Policy. (2022). Ensuring free, immediate, and equitable access to federally funded research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf

Plale, B., & Kouper, I. (2017). The centrality of data: Data lifecycle and data pipelines. In M. Chowdhury, A. Apon, & K. Dey (Eds.), Data Analytics for Intelligent Transportation Systems (pp. 91–111). Elsevier.

Resnik, D. B. (2020). What is ethics in research and why is it important? National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

U.S. Department of Education. (1974). Family Educational Rights and Privacy Act (FERPA).

U.S. Department of Health and Human Services. (2003). HIPAA privacy rule: Guidance regarding methods for de-identification of protected health information. https://www.hhs.gov/hipaa

U.S. Department of Health and Human Services. (2018). Federal policy for the protection of human subjects (Common Rule). https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones." In Proceedings of the ACM Conference on Ubiquitous Computing. 2014.