WIP: Advancing Data Quality Assurance and Privacy Protection Techniques: Bridging Theory and Practice

Amin Malek, SM IEEE
Department of Computer and Electrical Engineering
California State University, Bakersfield, CA, USA
aminmalek_m@ieee.org

Alberto Cruze
Department of Computer and Electrical Engineering
California State University, Bakersfield, CA, USA
acruz37@csub.edu

Norma Felix
Department of Computer and Electrical Engineering
California State University, Bakersfield, CA, USA
nfelix3@csub.edu

Erin Bangloy
Department of Computer and Electrical Engineering
California State University, Bakersfield, CA, USA
eviray@csub.edu

Abstract— **This work-in-progress (WIP) research-to-practice paper describes a work in progress by the authors to integrate appreciation of privacy, ethics, regulatory compliance, and research into Senior Project capstone experiences for Electrical and Computer Engineering. The student work focused on data quality assurance and de-identification topics to enhance quality, accuracy, completeness, consistency, and timeliness. Real-world data protection regulations grounded projects to meet ABET EAC Criterion 3 requirements for Student Outcome 2. Students explored the topics in a Project-Based Learning (PBL) format as a part of their senior project. In addition to implementing PBL, our focus for the senior project capstone is securing as many industrially sponsored projects as possible. This paper focuses on a few senior projects that are PBL, sponsored by industry, and emphasize data quality assurance and privacy protection techniques. We present a framework that meets assessment needs and uses project-based learning on a current topic of interest. The student findings offer insights into the theoretical and practical challenges and opportunities of implementing data quality assurance and de-identification techniques across different domains.**

Keywords: ABET assessment, project-based learning, inquiry-based learning, data quality assurance, de-identification

I . INTRODUCTION, BACKGROUND AND RELATED WORKS

A. Project-Based Learning

Project-based learning (PBL) originated in the early 20th century, primarily influenced by John Dewey's progressive education theories. [1]. Dewey advocated for an educational framework emphasizing learning through doing, which led to the development of experiential learning models, including PBL. This approach gained substantial traction in the 1970s with the advent of more student-centered learning philosophies that focused on critical thinking and problem-solving skills in real-world contexts.

PBL is a teaching method in which students gain knowledge and skills by working for an extended period to investigate and respond to a complex question, problem, or challenge. This approach emphasizes active learning by exploring real-world challenges and problems, encouraging deeper engagement than traditional instruction. In PBL, students typically collaborate in groups, make their own decisions, and engage in self-guided inquiry, culminating in a project that integrates and demonstrates their learning. Despite being decades old, many instructors continue to find ways to innovate curriculum through PBL [2, 3].

B. ABET Student Outcome 2

The Accreditation Board for Engineering and Technology (ABET) accredits post-secondary education programs in applied and natural science, computing, engineering, and engineering technology. Specifically, the Engineering Accreditation Commission (EAC) is one of ABET's commissions that accredits engineering programs. EAC requires that programs establish and document clear educational objectives consistent with the institution's mission and the needs of the program's various stakeholders.

This study centers on EAC Criterion 3, specifically Student Outcome 2 (SO2), which assesses a student's capability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors.

At California State University, Bakersfield, this criterion is primarily evaluated through the Senior Project capstone experience. Given the complexity of these interconnected issues, this paper explores PBL activities designed to assess these global and societal concerns, including their economic ramifications.

C. Motivation

Our research focuses on three questions about applying PBL in assessing Criterion 3 SO2 according to ABET standards and incorporating research on big data security. These questions are:

- How can PBL enhance the quality of the Senior Project capstone experience?
- How can the broad concepts of SO2 be seamlessly integrated into a single class or activity?
- What innovative research opportunities in big data, wireless communications, and security can students pursue?

As shown in Figure 1, we proposed a method to incorporate PBL into industrial-sponsored senior project experiences. Traditionally, as shown in Figure 2, students choose a project and develop a prototype or simulation of a product.

PBL, a student-centered educational strategy, involves students solving authentic problems through a comprehensive, interdisciplinary inquiry. While many senior projects or capstone experiences have incorporated elements of PBL for decades, we aim to emphasize inquiry-based learning, a fundamental aspect of PBL. Unlike typical senior project activities, which tend to be teacher-centered, our approach fosters inquiry-based learning. This method enhances the course quality, meets assessment requirements for SO2, and encourages students to explore new research in areas such as wireless security and big data.

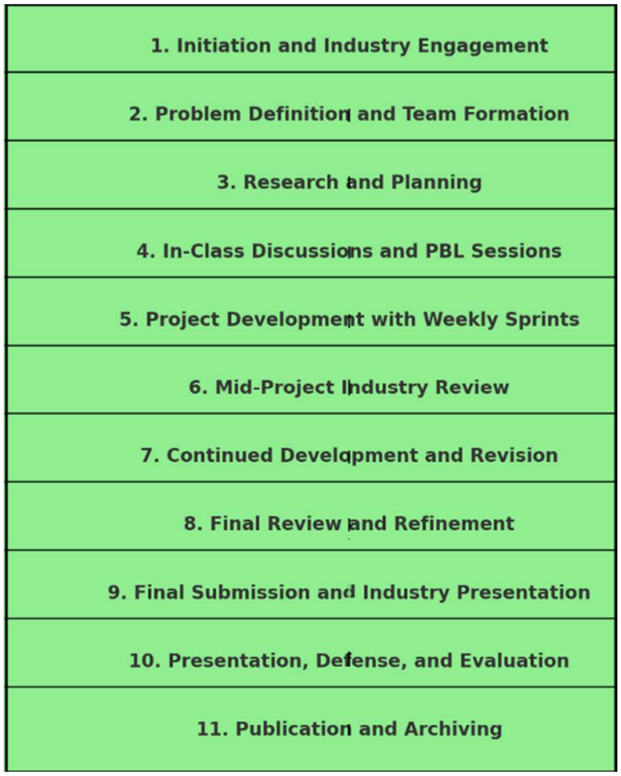

1. Initiation and Industry Engagement
2. Problem Definition and Team Formation
3. Research and Planning
4. In-Class Discussions and PBL Sessions
5. Project Development with Weekly Sprints
6. Mid-Project Industry Review
7. Continued Development and Revision
8. Final Review and Refinement
9. Final Submission and Industry Presentation
10. Presentation, Defense, and Evaluation
11. Publication and Archiving

Figure 1. PBL and industry Collaboration-based senior projects flowchart

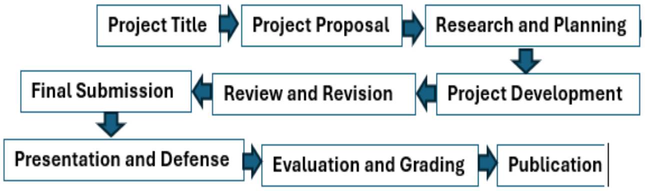


Figure 2. Traditional steps of the capstone project

II. Methods

A. Instructional Methods

The senior project capstone experience at the university spans a year and is divided into two unit sequences in the fall and spring semesters. A single instructor mentors the students throughout this period. Typically, students collaborate with a faculty supervisor, conduct literature reviews, and develop a significant project. Team collaboration is mandatory. The course emphasizes problem identification, analysis, and the application of engineering knowledge to propose solutions. The project must demonstrate intellectual merit based on the literature review. Key milestones are outlined as follows:

- The project must identify a critical area of electronics and big data from a literature review.
- Early completion of a design specification document that outlines concerns to public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors (SO2).
- Students give formal presentations in class to discuss progress. While students lead these activities, the format remains teacher-centric, with students primarily listening to instructor feedback.
- The instructor acts as the mentor, especially evident in the passive process of soliciting feedback and approval after presentations.
- Weekly sub-goals align with exploring the impact of security and big data concepts on public safety and economic factors, focusing on areas as challenges arise.
- Written assignments include both technical documents and non-technical essays, such as reflective pieces.

Assessment is primarily based on presentations and written activities, with the final project presentation serving as a critical deliverable. While a complete product is expected, it only constitutes some of the grade. The following modifications are implemented to enhance the inquiry-based learning experience for students:

- The project begins with a vague question or challenge related to the state of the art.
- Students must engage industry partners to ensure the project's authenticity.
- A planning document, more akin to a literature search than a market survey or software requirements document, is required.
- In-class discussions led by students on design and implementation foster active participation and promote inquiry-based learning, aligning PBL principles rather than

teacher-centric approaches.

- Weekly sprints replace predefined sub-goals, allowing for flexible adjustments as students progress.

B. Student Topics

In previous years, our primary focus was on digital communications, mainly novel modulation and multiplexing techniques, improving the available research works [4-7]. However, this year, our focus was changed to Data Engineering. The students' research focused on three critical areas: Data Quality Assurance (DQA), dimensions of data quality, and de-identification techniques. These topics were chosen due to their pivotal roles in ensuring the integrity and privacy of data, particularly in the context of big data.

DQA is essential for ensuring data is accurate, reliable, and fit for its intended purpose. By conducting a literature survey on big data packets, students explored various DQA techniques designed to detect, correct, and prevent errors in data. Data quality significantly impacts the reliability and validity of derived insights and outcomes in data-driven decision-making. With the advent of big data, characterized by its sheer volume and speed, traditional DQA methods are often inadequate. Therefore, techniques such as machine learning and artificial intelligence have become increasingly important. These advanced methods, including anomaly detection and predictive modeling, automate DQA processes, ensuring data integrity and supporting informed decision-making. The relevance of DQA in handling the complexities of big data makes it a crucial area of study for students aiming to contribute to organizational success through robust data management practices.

Dimensions of Data Quality encompass various attributes that define data quality, such as accuracy, completeness, consistency, and timeliness. Understanding these dimensions allows students to assess and improve data quality across different applications comprehensively. Each dimension plays a specific role in maintaining the overall integrity of the data, thereby ensuring that it can be trusted for critical business and analytical processes.

De-identification Techniques are vital for protecting privacy, especially when dealing with sensitive data such as personal information, such as ID numbers or car plate numbers. By learning and applying these techniques, students can effectively anonymize data, safeguarding individual privacy while maintaining the data's utility for analysis. Techniques such as masking, pseudonymization, and encryption are essential in complying with privacy regulations and preventing unauthorized access to personal information.

By delving into these interconnected topics, students not only enhance their understanding of data management but also contribute to the broader goal of ensuring data integrity and privacy in an increasingly data-driven world. This holistic approach to studying DQA, dimensions of data quality, and de-identification techniques equips students with the necessary skills to address contemporary challenges in data management and privacy protection.

C. Dimensions of Data Quality

Data quality dimensions encompass various aspects that contribute to data's overall reliability and usefulness. These dimensions include accuracy, completeness, consistency, and timeliness. Accuracy refers to the degree to which data reflects the true values or attributes. The framework of DQA using machine learning [8] highlights the importance of accuracy assessment in evaluating data quality, ensuring that data analysis and decision-making processes are based on reliable information.

Completeness denotes the extent to which data contains all the necessary attributes or information required for its intended purpose. When discussing completeness as a critical dimension of data quality, the survey on quality assurance techniques for big data applications emphasizes the importance of ensuring that no essential data is missing [9]. Consistency refers to the absence of contradictions or discrepancies within the data. The study on optimizing quality assurance strategies advocates for consistency checks and procedures to identify and rectify inconsistencies, ensuring that data remains coherent and reliable across different sources and formats [9]. Timeliness pertains to the currency and relevance of data in relation to the time it is needed for decision-making. Discussing timeliness as a crucial aspect of data quality in the context of big data applications highlights the need for strategies to ensure that data is up-to-date and available when required [4].

C.1. Procedures for Ensuring Data Quality

Ensuring data quality is crucial for accurate analysis and decision-making. The following are procedures used by our students to ensure data quality: 1- Data Collection (Source Verification, Consistency Checks, and Data Format Standardization) 2- Data Cleaning (Duplicate Removal, Missing Data Handling, **and** Error Detection 3- Data Validation (Range Checks and Cross-Validation)

C.2. Techniques for De-identification

Privacy safety is also crucial in this digital era, mainly when dealing with touchy domains such as healthcare. De-identification is an essential mechanism for maintaining anonymity while making allowances for the secondary use of data. Our students use the following procedures to ensure data quality for their senior projects related to de-identification: Masking (partial and complete masking), Anonymization (randomization and pseudonymization), Aggregation (grouping and summary statistics), Encryption, Suppression, and Data Shuffling.

III. Regulatory Frameworks and Ethical Considerations

In current data governance and privacy discussions, laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) play a crucial role. The GDPR, which came into effect in 2018 in the European Union, sets out detailed rules for safeguarding personal information and the rights of people in the EU and European Economic Area (EEA). It requires organizations handling personal data to meet strict criteria, including transparency, consent, minimizing data, and being accountable. In the same way, the CCPA passed in California in 2018 and has been in effect since 2020, marking a significant change in data privacy laws in the United States. This law gives California residents extensive control over their personal data, such as the right to be informed about collected information, the option to refuse its sale, and the ability to ask for their data to be deleted. The scope of the CCPA encompasses a wide range of businesses, regardless of whether they are located in California, thereby increasing its influence on worldwide data management.

A. Implications of Regulations on Data Quality Assurance and De-Identification Techniques

GDPR and CCPA impact data quality assurance and de-identification methods, requiring businesses to ensure data accuracy, relevance, and security. These regulations mandate informing individuals about data usage and allowing data deletion requests. De-identification methods like masking, encryption, and tokenization are essential to protect privacy while maintaining data utility and adhering to regulatory standards to prevent re-identification [7-9].

IV. Ethical Considerations in Data Handling Practices

Research and innovation with data present ethical dilemmas related to privacy, consent, and data ownership. The conflict lies in using data for scientific progress while preserving individual privacy. Big Data raises access control, policy integration, and authorization management issues. Ethical challenges include the fair use of data, especially from vulnerable groups, requiring strong guidelines and stakeholder involvement to ensure compliance with ethical norms and societal values [10].

Consent, fairness, and minimizing harm are crucial for ethical data management and protecting individual rights. Obtaining informed consent is vital but challenging with big data. Ensuring fairness involves addressing biases in data and algorithms to prevent discrimination. Strong security measures are necessary to prevent data breaches and misuse. Prioritizing these values builds trust and promotes responsible data management that is aligned with societal norms.

V. Preliminary Results

This section presents the preliminary results of implementing one of our industrial-based PBL projects, “a smart car garage device”. We focus on its effectiveness in ensuring data quality assurance and privacy protection. We begin by detailing the technical aspects of the device's functionality and then discuss its performance in real-world scenarios.

A. Technical Implementation

The smart car garage device uses license plate recognition (LPR) technology and de-identification algorithms. Upon detecting a vehicle approaching the garage, the device utilizes LPR to capture the license plate information. This information is then compared against a database of authorized license plates. If a match is found, the garage door is automatically opened, granting access to the vehicle. The device employs de-identification techniques to anonymize the data before storage or transmission for license plates that are not recognized. This process involves replacing identifiable elements of the license plate with non-identifying placeholders, preserving the integrity of the dataset while safeguarding individual privacy.

B. Performance Evaluation and Discussion

Evaluating PBL and industry collaboration-based senior projects requires a comprehensive approach encompassing various qualitative and quantitative metrics. This study has 16 students working in groups of 2 on eight different projects. They have already completed Senior Project I (50% of the project), and by December 2024, they will complete the entire project. By the end of this year, we will be able to conduct a complete evaluation of the implementation of industry-based PBL projects. Our key project evaluation metrics are Project Quality and Innovation, Industry Collaboration, Problem-Based Learning (PBL) Implementation, Project Development Process, Presentation and Communication, Learning Outcomes and Student Growth, Impact and Sustainability, Stakeholder Satisfaction, and Innovation and Research Contributions.

Since this is a WIP paper, we haven’t thoroughly evaluated the effectiveness of our new industrial-based PBL senior project, but for example, to assess the performance of our smart car garage device (Fig. 3), we conducted a series of experiments in simulated and real-world environments. The metrics evaluated include license plate recognition accuracy, de-identification effectiveness, and overall system reliability. Our device demonstrated high accuracy in recognizing authorized license plates in various lighting. This reliability ensures that only authorized vehicles gain access to the garage, enhancing security and convenience for users. The de-identification process (Fig. 4) proved vigorous, effectively anonymizing unrecognized license plates without compromising data integrity. Through randomized replacement of identifiable characters, the device mitigates the risk of unauthorized access or misuse of personal information. The smart car garage device exhibited consistent performance and operational reliability throughout our testing. It processed incoming vehicles promptly, minimizing delays and ensuring seamless access control.

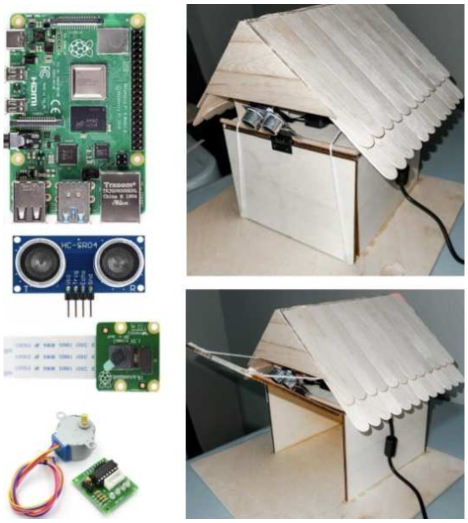
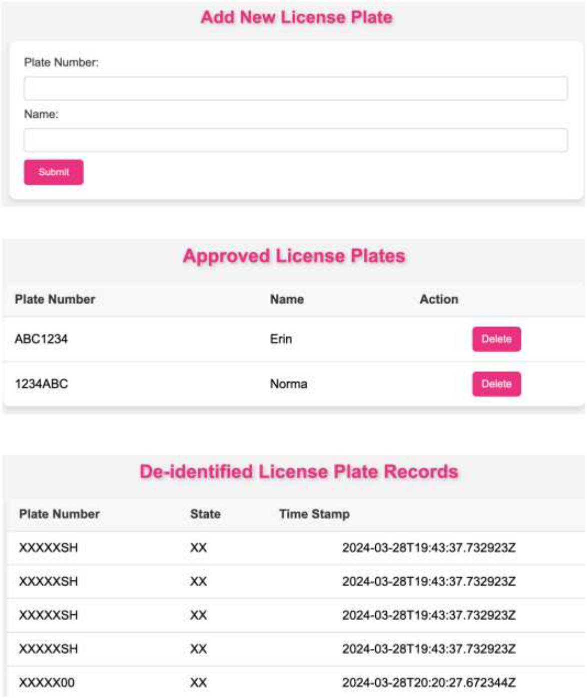


Figure 3. One of the senior design projects: Smart Car Garage device capable of recognizing and identifying plate numbers



Add New License Plate

Plate Number:

Name:

Approved License Plates

Plate Number	Name	Action
ABC1234	Erin	Delete
1234ABC	Norma	Delete

De-identified License Plate Records

Plate Number	State	Time Stamp
XXXXXSH	XX	2024-03-28T19:43:37.732923Z
XXXXXSH	XX	2024-03-28T19:43:37.732923Z
XXXXXSH	XX	2024-03-28T19:43:37.732923Z
XXXXXSH	XX	2024-03-28T19:43:37.732923Z
XXXXX00	XX	2024-03-28T20:20:27.672344Z

Figure 4. The outcome of our de-identification process

Our study's results demonstrate the feasibility and effectiveness of integrating data quality assurance and privacy protection measures into smart IoT devices such as smart car garages. By prioritizing technical robustness and ethical considerations, we contribute to developing a more equitable and trustworthy data ecosystem, fostering innovation while safeguarding individual privacy rights.

VI. Conclusion

In conclusion, this paper outlines ongoing efforts to incorporate privacy, ethics, regulatory compliance, and research considerations into PBL-based Senior Project capstone experiences in Electrical and Computer Engineering. Students engaged in PBL to delve into various data quality assurance and de-identification topics, aligning with ABET EAC Criterion 3 requirements. A key focus was striking a balance between privacy preservation and information utility. The framework presented fulfills assessment needs and facilitates PBL on contemporary topics. Student findings shed light on theoretical and practical aspects of implementing data quality assurance and de-identification techniques across diverse domains.

This paper also highlights the crucial need to uphold data accuracy and protect confidentiality in the age of big data and increased privacy reservations. By exploring different options for ensuring data quality and protecting privacy through de-identification methods, we have emphasized the importance of strong measures to improve data quality and uphold individuals' privacy rights. Students' research delves into the technical aspects of data management and the complex relationship between regulatory frameworks, ethical concerns, and practical obstacles encountered by global organizations. In addition, we have highlighted the changing regulatory landscape of laws like GDPR and CCPA, underscoring the need for flexible tactics to navigate legal challenges and maintain ethical principles. Moreover, our investigation into the ethical implications of using data for research and innovation highlights the importance of responsible data management and valuing the autonomy and well-being of individuals.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation under Grant No 2122442. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

References

[1] J. S. Krajcik and P. C. Blumenfeld, Project-based learning, Cambridge, 1975.

[2] P. Guo, N. Saab, L. S. Post, and W. Admiraal, "A review of project-based learning in higher education: Student outcomes and measures," International Journal of Educational Research, no. 102, p. 101586, 2020.

[3] M. A. Almulla, "The effectiveness of the project-based learning (PBL) approach as a way to engage students in learning," Sage Open, vol. 10, no. 3, 2020.

[4] N. Dong-Nhat, and A. Malekmohammadi. "Absolute added correlative coding: an enhanced M-PAM modulation format." Electronics Letters vol. 51, no. 20, pp. 1593-1595, 2015.

[5] R. Talib, M. F. Abdullah, A. Malekmohammadi, M.K. Abdullah, "Multi-slot and multi-level coding technique over amplitude-shift keying modulation for optical communication links. In 2011 16th European Conference on Networks and Optical Communications, pp. 161-164, 211

[6] M.A. Elsherif, and A. Malekmohammadi. "Power efficiency evaluation of mapping multiplexing technique and pulse amplitude modulation for noncoherent systems.", IEEE Photonics Journal, vol. 7 no. 4, pp. 1-11, 2015

[7] A. Malekmohammadi, G. A. Mahdiraji, M.K. Abdullah, A. F. Abas, A. Mokhtar, & M. F.A Rasid, "Absolute polar duty cycle division multiplexing technique." International Review of Electrical Engineering-IREE, vol 3, no 2, pp. 395-400, 2008

[8] N. B. Ding and E. Mit, "A framework of data quality assurance using machine learning," 13th International Conference on Information Technology in Asia (CITA), Aug. 2023. doi:10.1109/cita58204.2023.10262802

[9] P. Zhang, X. Zhou, W. Li, and J. Gao, "A survey on quality assurance techniques for Big Data Applications," IEEE Third International Conference on Big Data Computing Service and Applications, 2017

[10] F. Elberzhager and T. Bauer, "Optimizing Quality Assurance Strategies through an integrated quality assurance approach -- guiding quality assurance with assumptions and selection rules," in 40th EUROMICRO Conference on Software Engineering and Advanced Applications, Aug. 2014.

[11] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and D. Megias, "Individual differential privacy: A utility-preserving formulation of differential privacy guarantees," IEEE Transactions on Information Forensics and Security, vol. 12, no. 6, 2017, , pp. 1418–1429 doi:10.1109/tifs.2017.2663337

[12] A. R. Shovon, S. Roy, A. K. Shil, and T. Atik, "GDPR compliance: Implementation use cases for user data privacy in news media industry," in 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), May 2019. doi:10.1109/icasert.2019.8934660

[13] O. Amaral, S. Abualhaija, M. Sabetzadeh, and L. Briand, "A model-based conceptualization of requirements for compliance checking of data processing against GDPR," in IEEE 29th International Requirements Engineering Conference Workshops (REW), Sep. 2021. doi:10.1109/rew53955.2021.00009

[14] D. Singh, I. Nath, and P. K. Singh, "Security and privacy challenges in Big Data," Security, Privacy, and Forensics Issues in Big Data, 2020, pp. 97–124.
doi:10.4018/978-1-5225-9742-1.ch004