Teaching AI Through Narrative Data: A Practical Framework for Data Science and Retrieval-Augmented Generation

Charlie Dey
Texas Advanced Computing Center
charlie@tacc.utexas.edu

Susan Lindsey
Texas Advanced Computing Center
slindsey@tacc.utexas.edu

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) education has traditionally been split between technical model-building and data literacy. While these skills are often taught separately, the emergence of large language models (LLMs) offers an opportunity to unify them through narrative-driven, human-readable data transformation. This approach enables learners to query structured data using natural language while still engaging deeply with the underlying analytical processes.

We present a hands-on educational framework—debuting at the 2025 Big Data School in Costa Rica—that grounds AI learning in real-world data by transforming a single, richly structured dataset into narrative text that LLMs can ingest and reason over. Using the Austin Real-Time Traffic Incident Reports, participants apply core data science techniques—classification, clustering, regression, and forecasting—before extending their work into Retrieval-Augmented Generation (RAG) pipelines. A key design element is the careful selection of a dataset that supports multiple analytical tasks, contains time-series and categorical diversity, and can be effectively compressed into natural-language summaries without losing critical meaning. This continuity of dataset enhances comprehension, data intuition, and knowledge transfer without the cognitive effort of switching contexts.

By converting tabular data into concise, human-readable narratives, learners bridge traditional analytics and AI-enhanced insight generation. Outcomes include skill-building in Python-based ML pipelines, embedding and vector retrieval, and critical reflection on model interpretability, hallucination risk, and accessibility for non-technical users. This narrative-driven, single-dataset strategy—combined with explicit dataset selection principles—supports scalable adoption for high-performance computing and AI pedagogy in both academic and professional environments.

KEYWORDS

AI Education, Narrative Data, Retrieval-augmented Generation, Machine Learning Pedagogy, Data Storytelling

1 INTRODUCTION

Teaching artificial intelligence is no longer just about building models or tuning parameters — it's about helping people connect data, algorithms, and real-world meaning. In many AI and machine learning courses, these skills are taught in isolation: students learn data cleaning in one place, model building in another, and never see how the pieces fit together in practice. Large language models (LLMs) change that equation. By combining analytical rigor with the flexibility of natural language, they give us new ways to explore and explain data.

But there's a catch: most datasets used in teaching aren't built with this kind of integration in mind. They're either too small to show the complexity of real data, too large to work with interactively, or too fragmented to connect multiple techniques in a single flow. The result? Learners spend more time switching datasets and contexts than building a deep understanding of the data itself.

This work proposes a different approach: choose one rich, well-structured dataset and use it as the backbone for an entire AI learning journey. By converting that dataset into narrative form — short, human-readable stories that LLMs can reason over — we make it possible to bridge traditional analytics with LLM-powered exploration. The result is a smoother learning experience, where concepts build naturally on each other and learners see AI as both a set of tools and a way of thinking.

We present a framework for teaching AI that centers on a single, real-world dataset. Using the Austin Real-Time Traffic Incident Reports [2], participants explore the dataset through a series of ML tasks before transforming the data into narrative form for ingestion by LLMs. This supports a pedagogical model where learners move fluidly from traditional analytics to AI-enhanced question answering.

2 BACKGROUND AND RELATED WORK

AI and machine learning (AI/ML) education has expanded rapidly in recent years, driven by the growth of online courses, workshops, and bootcamps. While these initiatives have broadened access, they often share a common limitation: fragmentation. Learners are frequently exposed to isolated exercises tied to disparate datasets—a small CSV for cleaning in one session, a regression model on a different dataset in the next, and a visualization task on yet another. Each skill is valuable on its own, but this "patchwork" approach makes it difficult for students to see how the pieces of the data science pipeline—acquisition, cleaning, feature engineering, modeling, interpretation, and communication—fit together in practice.

Research underscores the importance of continuity in overcoming this barrier. Segel and Heer [7] emphasize that storytelling helps sustain engagement and comprehension in data-driven contexts, while Chen et al. [1] show that embedding narrative elements into data science courses strengthens the link between technical proficiency and conceptual understanding. Yet most AI/ML curricula

https://doi.org/10.22369/issn.2153-4136/17/1/7

remain disconnected, lacking a unifying dataset or sustained problem context that can scaffold learning across modules. As O'Reilly et al. [5] note, such fragmentation increases the cognitive burden on learners, who must continually reset their mental models as they move between unrelated domains and data structures.

This challenge is not unique to AI/ML. Huppenkothen et al. [4], in their work on hack weeks, highlight the benefits of sustained, project-based learning environments where participants engage with the full lifecycle of data-intensive research. Similarly, Powell et al. [6] demonstrate that hackathon-style approaches foster workforce development by immersing learners in end-to-end problem solving. Both approaches emphasize continuity, collaboration, and problem ownership—qualities often missing from short, fragmented workshop formats.

Our framework builds on these insights. By centering instruction around a single, complex dataset—the Austin Real-Time Traffic Incident Reports—we provide a continuous narrative arc that spans multiple instructional modules. This design allows students to progress from cleaning and feature engineering to classification, clustering, regression, and forecasting, all within the same context. Rather than treating algorithms as isolated tools, learners see how research questions, feature choices, and modeling decisions connect within a unified workflow that mirrors authentic AI/ML practice.

Narrative visualization has been recognized as a powerful means of communicating data and enhancing learner engagement [1, 7]. Embedding storytelling into analysis not only fosters clarity but also helps learners develop a deeper intuition for data. However, existing approaches rarely integrate ML training, narrative data transformation, and retrieval-based exploration into a single cohesive framework. Teaching with large datasets introduces further challenges of reproducibility, scalability, and accessibility [5]. Collaborative models like hack weeks [4] mitigate some of these issues, but few extend into modern AI workflows.

To situate our contribution within a broader pedagogical context, this framework aligns with principles of Project-Based Learning (PBL). PBL emphasizes sustained inquiry, real-world problem contexts, and iterative exploration, helping students develop both technical mastery and metacognitive skills. By anchoring the curriculum in a single, richly structured dataset, we extend PBL principles into the AI/ML domain: learners engage with the full lifecycle of data science, ask their own research questions, experiment with algorithmic choices, and reflect on successes and failures. This provides not only technical proficiency but also a holistic understanding of how data, methods, and narratives intersect in authentic practice.

3 METHODOLOGY

Our methodology is designed to create a cohesive, end-to-end learning experience that mirrors the lifecycle of a real AI/ML project. Rather than presenting algorithms in isolation, we anchor the workshop around a single, complex dataset that can sustain multiple analytical techniques and narrative perspectives. This design reduces the cognitive overhead of repeatedly switching between datasets, while allowing participants to progressively build intuition, technical proficiency, and storytelling skills in a unified context.

The methodology unfolds in four stages. First, we establish dataset selection principles to ensure that the chosen data can support diverse forms of analysis while remaining interpretable to learners. Next, participants engage in data familiarization, a stage focused on developing intuition about the dataset's structure, distributions, and potential features. Once a foundation of familiarity is established, the data is subjected to narrative transformation, where structured records are recast into natural language statements that remain semantically stable but more accessible for reasoning and retrieval tasks. Finally, participants integrate these narrative records into AI/ML and RAG workflows, combining traditional analytical techniques with retrieval-augmented generation pipelines.

Together, these stages create a scaffolded yet flexible framework that supports both technical mastery and narrative coherence, preparing learners to move fluidly between coding, analysis, and communication.

3.1 Dataset Selection Principles

Selecting the right dataset is central to the framework. In contrast to traditional workshops that rotate through small, disconnected datasets, we sought a dataset that could sustain **multiple analytical perspectives** and provide learners with opportunities to practice aligning research questions, features, and models. Our criteria for dataset selection were:

- **Enable multiple analytical techniques.** The dataset needed to naturally support a range of methods including classification, clustering, regression, and time-series forecasting. This diversity allows participants to experience how different models answer different types of questions, such as "What category does this belong to?" versus "What value will occur next?"
- **Contain temporal, categorical, and textual elements.** A variety of data types ensures that learners encounter the practical challenges of feature engineering, encoding, and preprocessing. For example, timestamps can drive forecasting models, categorical labels inform classification, and textual notes provide opportunities for embedding-based retrieval.
- **Balance complexity and accessibility.** The dataset must be large and realistic enough to feel "authentic," but not so large that it overwhelms learners during interactive exploration. This balance supports hands-on work while preventing the workshop from devolving into a purely computational challenge.
- **Be semantically stable under narrative transformation.** Because a key component of the framework is converting structured records into narrative form, the data had to retain its essential meaning whether represented as a table, a feature set, or a natural language statement.
- **Offer real-world relevance.** Context matters for engagement. A dataset tied to issues like urban traffic and public safety resonates with learners while also supporting authentic analytical challenges.

Beyond these structural considerations, the dataset also needed to facilitate **model selection as a pedagogical activity**. One workshop module asked participants to identify potential research questions and then map those questions to suitable models. For example, predicting whether an incident is likely to involve injury calls for a classification model, while estimating the expected response time benefits from regression. Clustering can be used to discover hidden patterns in incident types, and time-series forecasting enables projections of traffic incidents by time of day or season. By grounding these choices in a single, multifaceted dataset, students were able to see how different models carry distinct assumptions and data requirements, and why choosing "the right tool for the job" is a critical skill in AI/ML practice.

3.2 Data Familiarization

Before applying any machine learning techniques, participants engaged in a dedicated stage of **data familiarization**. This phase was designed to build what we call "data intuition"—an understanding of not only the structure of the dataset, but also its limitations, potential, and the kinds of questions it can reasonably answer.

Participants began by examining the dataset's schema, field definitions, and metadata. They explored distributions of numerical features, frequency counts of categorical variables, and patterns in temporal fields. For example, learners inspected how incidents varied by time of day, type of report, and location zone. This process was not simply descriptive but interpretive, as participants were encouraged to ask:

- What trends or anomalies stand out in the raw data?
- Which features seem relevant to the research questions we might ask?
- Are there missing values or inconsistencies that could bias the models?
- How might the temporal, categorical, or textual aspects of the data inform feature engineering?

Visual exploration played a central role. Using `pandas` for summary statistics and `matplotlib` for visualization, participants generated histograms, scatter plots, and heatmaps to observe relationships and outliers. These exercises helped learners become comfortable with the dataset before any modeling occurred.

The goal of this phase was twofold: (1) to reduce the cognitive burden of treating the dataset as a black box, and (2) to lay a foundation for later discussions about **model-data fit**. By first developing familiarity with the data, participants were better prepared to reason about which models would succeed, which would fail, and why. This aligns with project-based learning practices, where understanding the problem context is as important as technical implementation.

3.3 Narrative Transformation

Once participants had established a solid understanding of the dataset, the next step was to translate structured tabular records into short, human-readable narratives. This process, which we call **narrative transformation**, served two purposes: (1) to help learners consider how structured data can be expressed in natural language, and (2) to prepare the data for exploratory exercises with large language models (LLMs) later in the workshop.

For example, a traffic incident record originally stored as structured fields:

- Timestamp: 2025-08-14 17:43
- Location: Interstate 35, Zone 4
- Type: Collision
- Reporting Agency: Austin Police Department

could be transformed into a narrative record such as:

> "At 5:43 PM on August 14th, a collision occurred on Interstate 35 in zone 4 of Austin as reported by the Austin Police Department."

This conversion preserved the semantic content of the original record while making it accessible to both human readers and natural language systems. Participants experimented with different levels of detail—from minimal fact-based sentences to context-rich descriptions—and reflected on how narrative framing influences interpretation.

Importantly, the narrative transformation exercise was not just a prelude to LLM use. It also provided a bridge back to traditional feature engineering by raising questions such as:

- Which fields are essential to preserve, and which can be omitted without loss of meaning?
- How do phrasing and word choice affect interpretability for humans versus automated systems?
- What risks of ambiguity or bias are introduced during narrative construction?

Rather than immediately building a full retrieval-augmented generation (RAG) system, the workshop used these narrative records to *explore* how LLMs handle structured-to-text conversions, embedding creation, and simple retrieval tasks. This positioned narrative transformation as a conceptual hinge between conventional machine learning workflows and emerging AI-assisted data exploration, without requiring participants to fully master production-grade RAG pipelines.

3.4 Integrated AI/ML and Exploratory RAG Workflows

The final phase of the workshop was designed to connect traditional machine learning pipelines with exploratory uses of large language models (LLMs). This integration unfolded in two stages.

3.4.1 End-to-End AI/ML Pipelines. Participants first completed the full sequence of steps in a conventional machine learning workflow:

- **Data preparation** — Cleaning, normalizing, and splitting the traffic incident dataset into training and testing sets.
- **Feature engineering** — Identifying relevant features such as time of day, incident type, or location zone, and creating derived variables (e.g., weekday vs. weekend, rush-hour indicators).
- **Model exploration** — Matching research questions to algorithms, such as:
  - *Classification:* Can we predict whether an incident will require emergency response?
  - *Clustering:* What natural groupings of incidents emerge from location and severity?
  - *Regression:* How do weather or time features correlate with incident duration?
  - *Forecasting:* Can we project incident frequency trends across future weeks?

- **Model selection and tuning** — Comparing performance metrics, discussing overfitting risks, and iteratively refining hyperparameters.

This phase emphasized not only running models but also *choosing the right model for the right question.* By explicitly mapping data requirements (e.g., categorical vs. continuous features, time-dependence) to algorithm capabilities, students developed intuition for how methodological choices shape research outcomes [3].

3.4.2 Exploratory RAG Extensions. Once participants had completed the ML pipelines, they returned to their narrative-transformed records to explore how LLMs could add value. Rather than constructing a full production retrieval-augmented generation (RAG) pipeline, the workshop focused on lightweight experiments:

- Creating embeddings of narrative records and storing them in a vector database.
- Querying the vector database with natural language prompts (e.g., "Find recent collisions near Interstate 35 during rush hour").
- Comparing how well the retrieval results aligned with answers generated through traditional SQL or Pandas queries.
- Reflecting on issues of precision, ambiguity, and hallucination when using LLMs for data-grounded tasks.

This exploratory approach allowed students to see how narrative data transformations could be leveraged by modern AI tools, while avoiding the complexity of building full-scale RAG systems. In doing so, the workshop highlighted the complementary strengths of structured-query methods and LLM-based retrieval, preparing learners to critically assess when each is appropriate.

3.4.3 Hackathon-Style Application. To reinforce these skills, the workshop culminated in a hackathon-style exercise. Student teams selected their own datasets or worked with subsets of the traffic incident dataset, then:

- Defined three research questions of interest.
- Identified what features would be necessary to address those questions, including possible engineered variables.
- Mapped their questions to candidate algorithms, explicitly discussing which models could succeed or fail and why.
- Made assumptions about their data (e.g., independence, completeness, or stationarity) and presented these in team introductions.
- Implemented, tuned, and compared algorithms in practice.
- Presented their findings, reflecting on both successful and unsuccessful approaches.

This design encouraged participants to approach machine learning as an open-ended process of inquiry rather than a fixed recipe. The hackathon not only reinforced technical content but also fostered collaboration, problem framing, and critical thinking. We expand on the outcomes of this exercise in Section 3.5.

3.4.4 Exploratory RAG Extensions. Once participants had completed the ML pipelines, they returned to their narrative-transformed records to explore how LLMs could add value. Rather than constructing a full production retrieval-augmented generation (RAG) pipeline, the workshop focused on lightweight experiments:

- Creating embeddings of narrative records and storing them in a vector database.
- Querying the vector database with natural language prompts (e.g., "Find recent collisions near Interstate 35 during rush hour").
- Comparing how well the retrieval results aligned with answers generated through traditional SQL or Pandas queries.
- Reflecting on issues of precision, ambiguity, and hallucination when using LLMs for data-grounded tasks.

This exploratory approach allowed students to see how narrative data transformations could be leveraged by modern AI tools, while avoiding the complexity of building full-scale RAG systems. In doing so, the workshop highlighted the complementary strengths of structured-query methods and LLM-based retrieval, preparing learners to critically assess when each is appropriate.

3.5 Hackathon Outcomes

The hackathon-style exercise served as the culminating activity of the workshop, providing participants with an opportunity to synthesize technical content, methodological reasoning, and collaborative problem-solving. While prior sessions emphasized guided exploration of specific techniques, the hackathon placed students in a semi-open environment where they were responsible for framing their own research questions, selecting analytical approaches, and presenting results.

3.5.1 Team Problem Framing. Each team began by selecting a dataset (either a subset of the traffic incident data or an alternative dataset of personal interest) and proposing three guiding research questions. Examples included:

- *Classification:* Can we predict whether an incident requires a police response based on time, location, and type?
- *Regression:* What factors are most predictive of incident duration?
- *Forecasting:* Can we estimate how traffic incidents will evolve during upcoming weeks?

Teams were required to articulate the assumptions underlying their questions (e.g., data completeness, independence of samples, or stationarity in time series). This step highlighted the importance of transparency in computational science, as assumptions shape both the feasibility and interpretation of results.

3.5.2 Model Development and Experimentation. Once questions were defined, teams mapped them to candidate machine learning methods. For instance, classification problems often involved decision trees or logistic regression, while clustering tasks employed k-means or hierarchical approaches. Teams were encouraged to experiment with multiple algorithms, compare performance, and reflect on trade-offs such as interpretability versus predictive accuracy.

By aligning models to questions, participants gained a deeper appreciation of the iterative nature of AI workflows: each decision about features, algorithms, and tuning parameters reflected a hypothesis about the underlying data.

3.5.3 Presentations and Peer Review. Hackathon teams concluded with short presentations to the broader workshop group. Presentations included:

- A clear statement of research questions and motivating rationale.
- The dataset and features selected, including engineered variables.
- The algorithms applied, performance comparisons, and tuning efforts.
- A discussion of assumptions and potential limitations.
- Reflections on what worked, what failed, and how results could inform further inquiry.

Peer participants were invited to ask questions, challenge assumptions, and suggest alternative modeling strategies. This peer review element emphasized scientific communication and constructive critique as essential components of computational practice.

3.5.4 Observed Educational Impact. The hackathon outcomes demonstrated several key gains for participants:

- **Critical thinking:** Students learned to evaluate not only how models perform, but also whether they are appropriate for the question at hand.
- **Collaboration:** Teams reported that framing problems and dividing tasks required negotiation and clear communication.
- **Adaptability:** Many groups encountered data limitations or unexpected model behaviors, requiring them to pivot strategies.
- **Confidence:** By presenting to peers, participants demonstrated increased comfort discussing both successes and failures in technical work.

Collectively, these outcomes illustrate how hackathon-style integration can transform abstract machine learning concepts into lived experience. Rather than following step-by-step tutorials, participants engaged with the uncertainties and trade-offs that define real-world computational research.

4 OUTCOMES

The outcomes of this workshop can be considered along three dimensions: participant learning gains, the development of an integrated RAG-enabled workflow, and the collaborative hackathon that culminated the experience. Together, these outcomes provide evidence that the framework effectively bridges fragmented AI/ML training with a unified, project-based learning experience.

4.1 Learning Gains

Participants demonstrated measurable improvement in their ability to connect individual analytical tasks to the broader data science lifecycle. Early exercises revealed that many students could apply standard ML algorithms (e.g., linear regression or k-means) in isolation, but struggled to articulate how such methods mapped to different kinds of research questions. By the end of the workshop, participants showed stronger competency in aligning questions with appropriate methods, understanding data requirements for specific models, and communicating results in both technical and narrative forms. Informal assessments and group discussions suggested that this progression was due to the continuity of working with a single dataset throughout the workshop.

4.2 Integrated RAG Workflow

Another important outcome was the successful incorporation of a retrieval-augmented generation (RAG) pipeline. Instead of presenting RAG as an isolated "add-on," the workshop positioned it as a natural extension of the narrative transformation process described in Section 3. This allowed participants to experience how large language models (LLMs) can query, summarize, and contextualize data when supported by an external knowledge base. Embedding RAG within the workflow highlighted both its potential and its limitations: students noted that while the approach excelled in exploratory analysis and generating human-readable summaries, it required careful tuning and filtering to avoid hallucinations or irrelevant responses. These reflections helped frame RAG as a complementary, not replacement, technology within applied ML.

4.3 Collaborative Hackathon

The capstone hackathon provided the most visible demonstration of learning outcomes. In small teams, participants were tasked with selecting a subset of the traffic dataset and framing three distinct research questions. Teams had to justify why their chosen questions were both meaningful and feasible, identify the features necessary to answer them, and propose algorithms that could address those questions effectively. This process required teams to articulate assumptions and trade-offs, mirroring authentic data science practice.

During the hackathon, participants iteratively refined their approaches by cleaning data, engineering features, and experimenting with supervised and unsupervised models. Teams also reviewed their assumptions when the results did not align with expectations, reinforcing the importance of critical evaluation in ML workflows. The final presentations included not only algorithmic results but also narrative justifications of the methodological choices. This underscored the dual focus of the workshop: technical competence and the ability to communicate insights clearly. Feedback from participants indicated that the hackathon was the most engaging part of the workshop, providing a strong sense of ownership and achievement.

4.4 Collaborative Hackathon (Expanded)

The capstone hackathon provided a structured yet open-ended environment for participants to apply the skills they had acquired throughout the workshop. Each team was responsible for selecting a subset of the traffic dataset and developing three research questions that were both analytically tractable and pedagogically meaningful. Teams needed to evaluate what features would be required to address their questions effectively and determine which algorithms would best answer them. This process involved considerations such as data sparsity, feature types, temporal dependencies, and model assumptions, reflecting the types of decisions data scientists make in professional contexts.

Teams then presented their preliminary approaches, assumptions, and anticipated challenges during a "team introduction" session. This encouraged peer feedback and fostered a culture of collaborative problem solving. Throughout the hackathon, participants iteratively refined their models, conducted feature engineering, and compared performance across algorithms. Failures were treated as learning opportunities: when models underperformed or produced unexpected results, students revisited assumptions, discussed potential data issues, and considered alternative analytical strategies.

Final team presentations emphasized both the technical outcomes and the narrative framing of their work. Students demonstrated competence in applying classification algorithms, clustering methods, regression models, and forecasting approaches within a single dataset, illustrating how the workshop's design facilitated knowledge transfer across techniques. The hackathon experience highlighted critical aspects of AI/ML education: linking questions to methods, evaluating model suitability, and communicating findings effectively.

4.5 Limitations and Reflections

While the single-dataset approach proved effective for building continuity and cohesion, it also imposed certain constraints. The choice of dataset determined the scope of research questions and analytical techniques that could be explored. Some participants noted that a more diverse set of datasets might allow exploration of additional model types or data modalities. Additionally, time constraints limited the depth to which participants could explore complex models or advanced feature engineering techniques.

Despite these limitations, feedback suggested that the integrated, project-based format enhanced participant confidence and engagement. The emphasis on asking the right questions, mapping them to appropriate algorithms, and critically reflecting on model assumptions provided a holistic understanding of the AI/ML pipeline. Participants also reported increased appreciation for the challenges inherent to real-world data and the value of iterative, exploratory learning in mastering AI/ML workflows. By structuring the workshop around a single, richly detailed dataset, learners experienced continuity across activities that is rarely seen in traditional AI/ML trainings. The Project-Based Learning (PBL) approach allowed participants to formulate research questions, explore feature requirements, and select models appropriate to the data and questions at hand. As illustrated in Table **??**, each activity built on the previous one, reinforcing key skills and concepts while maintaining a coherent narrative thread. This continuity not only reduced cognitive load typically associated with switching datasets and contexts but also enabled students to observe how decisions in data preprocessing, model selection, and evaluation interconnect in real-world workflows. The hackathon portion further strengthened these outcomes by letting learners independently apply the same pipeline on novel datasets, fostering critical thinking, problem-solving, and reflection on algorithm suitability and data quality.

5 NOVELTY AND IMPACT

This structured, continuous learning experience not only helped participants gain technical skills but also reinforced critical thinking and problem-solving across the full AI/ML workflow. By guiding learners from data familiarization through feature engineering, model selection, and evaluation within a single dataset, the workshop fostered a deeper understanding of algorithmic choices and data constraints. These outcomes naturally lead into the broader discussion of the pedagogical novelty and impact of this approach, highlighting how integrating project-based learning with a continuous dataset framework differentiates this workshop from more fragmented AI/ML training models. This workshop framework differs from most AI/ML educational offerings in several key ways:

- **Single-dataset integration across the curriculum:** By centering instruction on a single, rich dataset—the Austin Real-Time Traffic Incident Reports—participants experienced continuity across multiple analytical tasks, from data cleaning to feature engineering, classification, clustering, regression, and forecasting. This approach reduced the cognitive effort typically associated with switching datasets, helping learners build deeper data intuition and a cohesive understanding of the AI/ML pipeline.
- **Project-based, question-driven learning:** Incorporating a capstone hackathon emphasized asking the right questions, selecting appropriate models, and iteratively refining solutions. This mirrors real-world data science workflows and leverages learning through failure, encouraging critical thinking, problem ownership, and collaborative skill-building.
- **Bridging analytical techniques with model selection principles:** Students were explicitly guided to consider which algorithms were best suited to their research questions and dataset characteristics. This included discussions on data criteria (e.g., feature types, sparsity, temporal dependencies), algorithm assumptions, and the interpretability of model outputs—topics often overlooked in conventional workshops.
- **Reproducibility and scalability:** All code and workflows were implemented in Python using open-source libraries, ensuring that the framework can be adapted for other datasets, institutions, or professional settings. By providing detailed instructional materials, the workshop design supports reproducible and scalable AI/ML pedagogy.
- **Holistic pedagogical impact:** Unlike traditional workshops that isolate skills, this integrated approach helps learners develop a broad understanding of the end-to-end AI/ML process. Participants reported increased confidence in applying methods across different tasks, linking questions to algorithms, and communicating findings—skills directly relevant to both academic and industry contexts.

Taken together, these elements contribute to a novel educational paradigm that emphasizes continuity, practical skill-building, and reflective learning. By combining project-based instruction with single-dataset integration, this framework addresses a documented gap in AI/ML education: the fragmentation of learning experiences across multiple, disconnected datasets and exercises. This paper provides a model that other educators can replicate, adapt, and expand upon in diverse instructional settings.

Table 1: Workshop Outcomes: Learning Objectives, Activities, and Skills Developed

Learning Objective	Workshop Activity	Skills Developed
Data Familiarization	Exploring the Austin Traffic Incident dataset, examining distributions, spotting anomalies, visualizations	Data intuition, EDA, feature identification
Data Cleaning and Feature Engineering	Handling missing values, encoding categorical variables, feature scaling	Python/Pandas, preprocessing pipelines, reproducible workflows
Classification	Logistic regression, decision trees; discussion of entropy, information gain, CountVectorizer/TfidfVectorizer	Model understanding, feature importance, algorithm selection, analytical reasoning
Clustering	K-Means, DBSCAN; evaluating cluster quality	Unsupervised learning, parameter tuning, pattern discovery
Regression	Linear regression, Random Forest, revisiting overfitting/underfitting	Predictive modeling, model evaluation, error analysis
Time-Series Forecasting	ARIMA; trend and seasonality analysis	Forecasting, temporal reasoning, model selection criteria
Hackathon Project	Teams choose a dataset, formulate 3 research questions, determine needed features and algorithms, run experiments, present results	Problem formulation, dataset-to-question mapping, experimental design, teamwork, presentation skills

6 DISCUSSION AND FUTURE WORK

The workshop described in this paper demonstrates a practical, integrated approach to AI/ML education, yet there are several areas for refinement and expansion. One key limitation is the scope of assessment: while qualitative feedback and student reflections indicate enhanced comprehension and confidence, future iterations will incorporate structured evaluation metrics to measure learning gains, algorithmic understanding, and data literacy improvements. For example, pre- and post-workshop assessments, alongside rubrics for hackathon projects, could provide quantitative evidence of skill acquisition.

Reproducibility and scalability are also central considerations. All materials were developed using open-source Python libraries and documented workflows, enabling replication at other institutions. However, wider deployment may require adaptation for different datasets, class sizes, or learner backgrounds. Future work will explore modularizing the framework so that instructors can plug in alternative datasets while maintaining continuity and coherence across tasks.

Expanding the workshop series will also provide the opportunity to collect longitudinal data on learning outcomes, model selection proficiency, and project-based reasoning. Tracking how learners transfer skills to new datasets or real-world problems will help validate the framework's effectiveness beyond the initial pilot. Additionally, integrating more explicit discussions of ethical considerations, bias in models, and interpretability trade-offs will strengthen students' understanding of responsible AI practice.

Finally, future iterations could examine hybrid or fully remote implementations, leveraging cloud-based platforms or high-performance computing resources to broaden accessibility. By iterating on these dimensions, we aim to create a replicable, evidence-based model for comprehensive AI/ML instruction that addresses fragmentation, enhances learner engagement, and prepares participants for authentic data-driven challenges.

ACKNOWLEDGEMENTS

We thank the organizers of the 2025 Big Data School in Costa Rica for the opportunity to pilot this workshop. We also thank fellow TACCsters Dr. Joe Stubbs and Dr. Anagha Jamthe for their work in developing an extensive course curriculum on AI/ML for the Computational Engineering department at the University of Texas, Austin [8].

REFERENCES

[1] Lujie Karen Chen, Jamie Gillan, Matthew Decker, Egan Eteffa, Anjelica Marzan, Justin Thai, and Sarah Jewett. 2023. Embedding Digital Data Storytelling in Introductory Data Science Course. *Journal of Problem Based Learning in Higher Education* 11, 2 (2023).

[2] City of Austin Open Data. 2025. Real-Time Traffic Incident Reports. https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x. Accessed: 2025-9.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.

[4] Daniela Huppenkothen, Anthony Arendt, David W. Hogg, Karthik Ram, Jacob T. VanderPlas, and Ariel Rokem. 2017. Hack weeks as a model for data science education and collaboration. *ArXiv preprint arXiv:1711.00028* (2017).

[5] Catherine M. O'Reilly, Tanya Josek, Rebekka D. Darner, and Sarah K. Fortner. 2022. Pedagogy of teaching with large datasets. *Biochemistry and Molecular Biology Education* 50, 5 (2022), 439–447.

[6] Jeaime Powell, Linda Bailey Hayden, Amy Cannon, John Holly, Charlie Dey, and Alexander Nolte. 2022. The HackHPC Model: Fostering workforce development in high-performance computing through hackathons. *Gateways 2022 Proceedings*. https://par.nsf.gov/biblio/10450534

[7] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148.

[8] Joe Stubbs and Anagha Jamthe. 2025. COE 379L: Software Design for Responsible Intelligent Systems, Spring 2025. https://coe379l-sp25.readthedocs.io/en/latest/. Course materials and syllabus for a machine learning and data systems class at The University of Texas at Austin.