CASTCurate: An Agentic System to Accelerate the Collection and Annotation of Data-Driven Stories

Aswin Janakiraman
University of Maryland, Baltimore County
Baltimore, MD, United States
aswinkj1@umbc.edu

Taha Hassan
The University of Alabama
Tuscaloosa, AL, United States
thassan1@ua.edu

Shenglin Li
The University of Alabama
Tuscaloosa, AL, United States
sli90@ua.edu

Lujie Chen
University of Maryland, Baltimore County
Baltimore, MD, United States
lujiec@umbc.edu

Jiaqi Gong
The University of Alabama
Tuscaloosa, AL, United States
jiaqi.gong@ua.edu

Abstract

This study introduces an AI-powered data storytelling agent designed to support data science educators by automatically curating high-quality, real-world data stories. The system streamlines the discovery of relevant instructional examples for specific teaching activities, including assignments, quizzes, classroom discussions, and case studies, by utilizing automated classification and narrative analysis. Our prototype significantly reduces instructor preparation time while improving the diversity, quality, and pedagogical alignment of curated stories. This innovation enables educators to more efficiently source, annotate, and deploy impactful data narratives tailored to their teaching and research objectives.

CCS Concepts

• **Information systems → Web searching and information discovery**.

Keywords

Data storytelling, data science education, agentic system

ACM Reference Format:
Aswin Janakiraman, Taha Hassan, Shenglin Li, Lujie Chen, and Jiaqi Gong. 2026. CASTCurate: An Agentic System to Accelerate the Collection and Annotation of Data-Driven Stories. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2 (SIGCSE TS 2026), February 18–21, 2026, St. Louis, MO, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3770761.3777345

1 Introduction

Data storytelling is an emerging domain with utility in diverse fields, including journalism, business communication, and bioinformatics [6]. However, incorporating data storytelling in computing and data science curricula poses unique challenges. One of these challenges is the selection of high-quality data-driven stories suitable for instructional use. Identifying data stories is complex, nuanced, and time-intensive. The quality of discovered examples can be inconsistent; they frequently lack statistical rigor, effective data visualization, or a clear narrative structure. Educators also struggle to find diverse examples that span critical domains, such as economics, health, and technology. The core difficulty lies in identifying stories that align precisely with specific pedagogical objectives and narrative frameworks required for varied educational contexts, including assignments, quizzes, and case studies. These narrative frameworks are essential pedagogical tools that structure data to tell a specific kind of story, for instance, illustrating change over time (Time-Based Progression), providing an overview before drilling into details (Overview to Detail), explaining a causal link (Cause-and-Effect), or framing the narrative as an investigative question and answer (Question-and-Answer).

While automated research agents exist, they offer limited support for these specialized needs. Current systems, such as Langchain MCP and Microsoft Deep Research Agents, are designed for broad information gathering but lack the domain-specific filters and classification capabilities necessary to evaluate data stories for educational use. Consequently, there is currently no system specifically designed to identify, classify, and curate data-driven narratives for pedagogical purposes. To address this critical gap, this poster introduces an AI-powered data storytelling agentic system designed to support data science educators. By utilizing automated classification and narrative analysis, our system streamlines the discovery of high-quality data stories. This innovation significantly reduces instructor preparation time while improving the diversity, quality, and pedagogical alignment of curated instructional materials.

2 Related Work

Recent agentic AI advances, particularly Deep Research Agents (DR agents), enable automated gathering, reasoning, and reporting over complex, noisy data sources [3]. In software engineering, deep research agents support multi-step exploration and structured context gathering, outperforming baselines in understanding large codebases [7]. In education, generative AI agents and dashboards provide automated scaffolding and personalized feedback, improve engagement, and bring real-world data into curricula [5, 8]. Multi-agent generative systems further automate data-driven storytelling,

SIGCSE TS 2026, St. Louis, MO, USA

ACM ISBN 979-8-4007-2255-4/2026/02
https://doi.org/10.1145/3770761.3777345

synthesizing multimodal narratives and visualizations suitable for classroom use [1, 4]. However, existing solutions often fall short in educational data curation: they lack robust mechanisms for collecting legitimate, domain-relevant data stories, systematic frameworks for pedagogical classification, and efficient reduction in instructor effort. Our approach addresses these gaps by automatically sourcing credible data stories, optimizing classification for instructional purposes, and significantly reducing the time needed for educators to find, annotate, and deploy topical, example-rich materials, therefore supporting better explanations and enriched teaching.

3 System Design and Evaluation

The CASTCurate architecture (Figure 1) is built around three core agents that collaboratively automate the data story pipeline. The process begins with the Data Collection Agent, which acts as the entry point by receiving an instructor's topic query. This agent applies advanced prompt engineering, leveraging both instructional frameworks and a large language model (LLM) to generate nuanced, context-sensitive classification criteria tailored to the specific pedagogical objectives. Using these dynamically generated criteria, the Web Scraping Tools are then deployed to systematically search, extract, and capture relevant data stories and rich metadata from both mainstream and domain-specific online sources. This automated retrieval not only broadens the scope of materials but also accelerates discovery, far outpacing manual search methods. Once candidate stories are gathered, the classification agent takes over, applying established narrative frameworks such as problem solution, sequence, cause-effect, comparative analysis, and the classic data story arc. Through multi-level annotation and structured tagging, the Classification Agent ensures that each story is contextually relevant, pedagogically valuable, and explainable. Only those stories deemed of high quality and surpassing a stringent confidence threshold are curated for instructional use. A small pilot evaluation of CASTCurate output reveals the following opportunities and challenges:

- **Accuracy**: CASTCurate reduces narrative discovery and annotation times, but it can occasionally collect URLs featuring carousels of data stories rather than individual stories.
- **Quality**: 90+% of the stories curated by the system are judged as relevant and pedagogically aligned by human reviewers.
- **Explainability**: Metadata and multilevel annotations enhance transparency and adaptation for instructional use, a feature not widely matched by peer systems.
- **Scalability**: The system enables rapid scalability; users can obtain and classify an arbitrary number of data stories within minutes, accelerating the discovery process compared to manual collection and classification.

Overall, CASTCurate's robust pipeline empowers educators with domain-specific, high-confidence examples while supporting transparent curation and ease of instructional integration.

4 Implications for Educators and Future Work

CASTCurate promises to assist CS and data science educators in a multitude of ways. It can support the design of quizzes and projects by curating sample training datasets to inform AI/ML model development. Annotated data stories can be transformed into timely and compelling (1) case studies, (2) group discussion prompts, and (3) visualization challenges. In our future work, we plan to combine CASTCurate with a user-facing AI writing assistance tool and add LMS support [2] to create reusable critical thinking exercises that allow students to compare their chosen narrative structures with AI-picked narratives. We also plan to classify these stories by complexity and knowledge prerequisites in order to create personalized assignments and A/B tests.

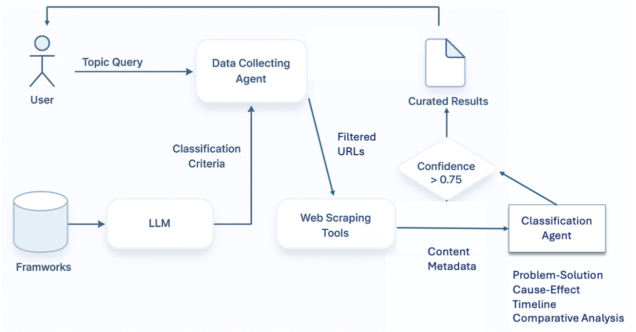

Figure 1: An overview of our proposed data story collection and annotation system.

5 Conclusion

CASTCurate bridges a critical gap in CS/DS educational technology by automating the collection and classification of data stories for instructional use. It streamlines narrative discovery and annotation, and delivers explainable examples with the potential to support diverse pedagogical goals. By enabling rapid scaling and reducing the manual search burdens on instructors, CASTCurate demonstrates clear advantages in efficiency, quality, and transparency over existing solutions. In our future work, we plan to craft high-resolution taxonomies for data stories, support instructor feedback, and improve the robustness of CASTCurate to noise and LLM updates.

References

[1] Samee Arif, Taimoor Arif, Muhammad Saad Haroon, Aamina Jamal Khan, Agha Ali Raza, and Awais Athar. 2024. The art of storytelling: Multi-agent generative ai for dynamic multimodal narratives. *arXiv preprint arXiv:2409.11261* (2024).

[2] Taha Hassan, Bob Edmison, Larry Cox, Matt Louvet, Daron Williams, and D Scott McCrickard. 2020. Depth of use: an empirical framework to help faculty gauge the relative impact of learning management system tools. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. 47–53.

[3] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, et al. 2025. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096* (2025).

[4] Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. DataNarrative: Automated data-driven storytelling with visualizations and texts. *arXiv preprint arXiv:2408.05346* (2024).

[5] Haotian Li, Lu Ying, Haidong Zhang, Yingcai Wu, Huamin Qu, and Yun Wang. 2023. Notable: On-the-fly assistant for data storytelling in computational notebooks. In *Proceedings of the 2023 Conference on Human Factors in Computing Systems*. 1–16.

[6] Kay Schröder, Wiebke Eberhardt, Poornima Belavadi, Batoul Ajdadilish, Nanette van Haften, Ed Overes, Taryn Brouns, and André Calero Valdez. 2023. Telling stories with data–A systematic review. *arXiv preprint arXiv:2312.01164* (2023).

[7] Ramneet Singh et al. 2025. Code Researcher: Deep Research Agent for Large Systems Code and Commit History. *arXiv preprint arXiv:2506.11060* (2025).

[8] Lixiang Yan, Roberto Martinez-Maldonado, Yueqiao Jin, Vanessa Echeverria, Mikaela Milesi, Jie Fan, Linxuan Zhao, Riordan Alfredo, Xinyu Li, and Dragan Gašević. 2025. The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics. *Computers & Education* (2025), 105322.