Visualization Retrieval for Data Literacy: Position Paper

Huyen N. Nguyen
huyen_nguyen@hms.harvard.edu
Harvard Medical School
Boston, Massachusetts, USA

Nils Gehlenborg
nils@hms.harvard.edu
Harvard Medical School
Boston, Massachusetts, USA

Abstract

Current resources for data literacy education, such as visualization galleries and datasets, provide useful examples but lack mechanisms for learners to query, compare, and navigate the visualization design space efficiently. This position paper advocates for visualization retrieval as essential infrastructure for data literacy, transforming static collections into dynamic, inquiry-based learning environments. We analyze the role of retrieval across the data lifecycle, demonstrating how it facilitates design space exploration and vocabulary expansion, supports data consumption through visualization comparison and critique, and aids data management via resource curation. We outline key opportunities for future research and system design, including integrated retrieval-authoring environments, pedagogical relevance modeling, and collaborative educational corpora. Ultimately, we argue that visualization retrieval systems empower learners to articulate intent, bridge technical barriers, and proactively reason with data.

CCS Concepts

• **Information systems** → **Information retrieval**; **Specialized information retrieval**; • **Human-centered computing** → **Visualization theory, concepts and paradigms**.

Keywords

Visualization Retrieval, Data Literacy, Design Space

ACM Reference Format:
Huyen N. Nguyen and Nils Gehlenborg. 2026. Visualization Retrieval for Data Literacy: Position Paper. In *CHI 2026 Workshop on Data Literacy, April 13, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages.

1 Introduction

Recent research in data science education emphasizes the importance of real-world examples for developing data literacy. Within this space, data visualization acts as both a motivational tool [18] and a means of communication [23]. The concept of *visualization literacy* has emerged as a distinct competency, encompassing the ability to read, interpret, and produce visual representations of data [4]. Although growing collections of visualization databases and datasets [3, 5, 7, 13, 15, 26] provide valuable pedagogical resources, these repositories remain largely static and curator-driven. Most importantly, the ability to retrieve specific, relevant examples is currently limited.

This lack of retrieval capability prevents students from navigating the design space more effectively. Prior work shows that exposure to relevant visualization examples helps participants understand which visual encodings and interactions are possible for their specific problems [21], and that visualization designers routinely search for visual examples as sources of inspiration [1]. To facilitate this, learners need mechanisms to ask questions of these collections, using either images or text keywords to drive their exploration. Aligning with the principles of inquiry-based learning [16], visualization retrieval offers a promising pathway for data literacy education. However, current systems lack the necessary support for exploring datasets through visual evidence, and ultimately navigating design spaces intentionally and proactively.

In this position paper, we argue that *visualization retrieval* is an essential, yet underexplored, infrastructure for data literacy. We address this gap by (1) contextualizing visualization retrieval within data literacy education, (2) mapping its role across the data lifecycle—production, consumption, and management, and (3) outlining opportunities for system design and research that integrate retrieval into teaching practice.

2 Background: Visualization Retrieval

Information retrieval systems are the infrastructure behind modern knowledge acquisition, yet the application of these principles to data visualization remains a specialized challenge. As defined in foundational information retrieval literature [19], the core objective is to find relevant documents that satisfy a user's information need through query mechanisms. In the context of visualization, this involves addressing the semantic gap: the challenge of translating user intent (e.g., *"I want to show a correlation"*) into low-level visual features (e.g., scatterplots, regression lines) that a system can index. For example, a novice student in genomics might type *"A T C G plot"* when they are actually looking for sequence logo visualizations, where each position in the DNA is shown as a stack of nucleotide letters A, T, C, and G, and the size of each letter reflects how often it appears. Without knowledge of the term *"sequence logo,"* they are unable to enter a precise query, as illustrated in the left panel of Figure 1.

In contrast to standard multimedia retrieval, visualization retrieval requires a deeper understanding of the data-binding structure within visual representations. The nature of visualization encompasses not just the image, but also domain tasks, user interaction, and human interpretation—which can be subjective and nuanced based on audience. Prior literature on visualization retrieval systems [6, 11, 17, 21, 22, 24, 27–29] has addressed these challenges by developing specialized similarity measures and query mechanisms, and generally follows a common retrieval loop from the user's perspective: (1) the user articulates a query conveying their intent, (2) the system matches and ranks candidates based on visual or semantic similarity, and (3) the user reviews and refines the search. As this iterative cycle mirrors the sensemaking process [12] and reinforces inquiry-based learning [16], we argue that it is well-suited for educational contexts. Through this process, retrieval supports not only access to examples but also the articulation of intent, a core component of data literacy and critical

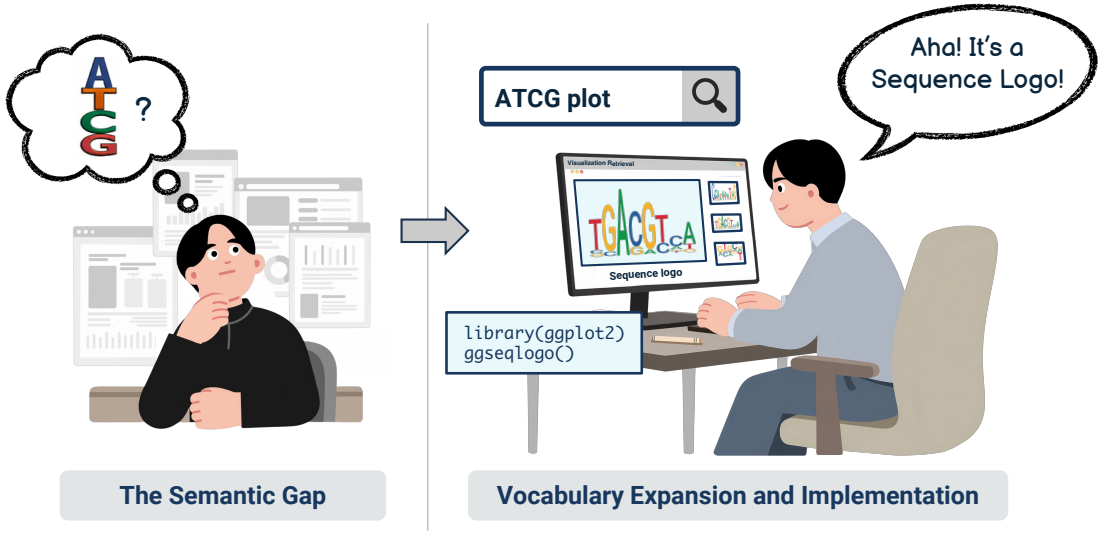

Figure 1: Illustration of visualization retrieval bridging the semantic gap. A student searching for an *"ATCG plot"* lacks the formal vocabulary to find *"sequence logo"* visualizations. A visualization retrieval system can match this informal query to sequence logo examples and return them with the correct terminology and associated code, helping the learner recognize the appropriate chart type, expand their visual vocabulary, and see how such a visualization can be implemented.

thinking with data. The next section continues with the implications of visualization retrieval for data literacy in detail.

3 The Role of Visualization Retrieval in Data Literacy

Adopting the framework of data literacy as a set of practices across the data lifecycle [14], we investigate the role and extent of visualization retrieval in data production, consumption, and management. Through this analysis, we highlight the relevance of visualization retrieval as essential literacy infrastructure. In this paper, we use the following working definitions:

(1) **Data production**: activities in which learners and practitioners create or modify visualizations to explore data or communicate findings. The learner is a producer or designer of visualizations.
(2) **Data consumption**: activities in which learners and practitioners encounter, interpret, critique, or rely on visualizations created by others. The learner is a reader, interpreter, or critic of visualizations.
(3) **Data management**: activities in which both learners and educators act as curators of datasets and visualizations, creating and maintaining repositories for future use.

3.1 Visualization Retrieval and Data Production

In production-related activities, students are responsible for *authoring* visualizations, including when they start from templates or existing examples. Visualization retrieval can support them by making the design space visible and easy to navigate, providing implementable examples, and expanding their visual vocabulary.

Design space exploration: In the early stage of visualization design, learners must answer questions involving design space such as: *"What visual encodings are suitable for my data (e.g., temporal, multivariate, spatial)?"* or *"How can I support the analytic task (e.g., comparison, trend detection, anomaly detection)?"* Searching for a certain visualization task or style, the system might return line charts, small multiples, stacked area charts besides the default of bar plots [27, 29]. In this case, a visualization retrieval system can act as a map of the design space, providing multiple *plausible designs* for the same underlying task. From this exploration, learners can compare trade-offs between options and thus make more intentional design decisions.

Inspiration and implementation: In practice, visualization authoring often follows a find-and-adapt workflow [2, 25]. Learners can search for an example that resembles their desired output, then adapt its code or specification to their own data and context. While static galleries like the R Graph Gallery [10] or Python Graph Gallery [9] provide code examples, they rely on manual browsing, forcing users to sift through the collections to find relevant templates. Visualization retrieval systems can significantly streamline this workflow by returning examples directly linked to code, specifications, or notebooks. By allowing learners to filter results by library (e.g., R, Python, React) or data structure, retrieval systems connect abstract design choices to their concrete actualization in code. This not only improves efficiency but reinforces data literacy by helping learners map visual encodings to the data transformations required to produce them.

Expanding visual vocabulary: Retrieving and viewing diverse charts also help expand the *visual vocabulary* for learners. Returning to the earlier genomics example, a learner may type *"ATCG plot"* and, through retrieval, encounter not only sequence logo visualization images, but also domain-specific naming such as *"sequence logo"* and *"sequence motif logo,"* as illustrated in the right panel of Figure 1.

3.2 Visualization Retrieval and Data Consumption

Data consumption refers to how individuals encounter, interpret, evaluate, and potentially act upon visualizations produced by others. In these activities, learners are not primarily responsible for creating visualizations; instead, they are asked to read, understand, critique, and compare existing visualizations. As understanding visual patterns and pitfalls is an essential component of visualization literacy [8], visualization retrieval supports this process by enabling intent-driven exploration of existing visualizations for comparison and critique.

With visualization retrieval, learners can not only explore the design space both with and without purpose [1], but also get to articulate their intents through refinements of searches. Systems such as VAID [29] and Geranium [21] support this type of inquiry through multimodal query methods ranging from categorical selection to multimodal inputs (e.g., free text, images, or declarative grammar). Consequently, while static galleries provide exposure to design varieties in a relatively passive manner, visualization retrieval offers a dynamic shift for learners to not only search but articulate their intents and refine the queries over time.

3.3 Visualization Retrieval and Data Management

Beyond production and consumption, retrieval aids in data management. As learners build their expertise, they need to organize examples. Retrieval systems can help learners curate personal libraries such as good practices or common pitfalls, and tag examples for future reference; this curation process itself reinforces literacy skills. On the other hand, from an educator's perspective, having a searchable knowledge base of examples helps streamline the development of curriculum materials. Opportunities in this area are presented in the next section.

4 Opportunities for Research and System Design

The integration of visualization retrieval into data literacy presents rich opportunities for research and system design. These directions are interlaced between data production, consumption, and management; we propose the following areas to support learners as they move across these practices, addressing both research and education purposes.

4.1 Integrated Retrieval-Authoring Systems

To fully streamline the process from finding inspiration to exploring the design space and to implementation, we call for development of integrated retrieval-authoring environments. While current workflows often require switching between a search engine (e.g., Google search) and a coding interface, future systems should bridge this gap by:

Embedding retrieval in authoring tools and vice-versa: Developing plugins for environments such as Jupyter or RStudio that retrieve adaptable code templates alongside the user's data, and providing previews on how the resulting visualization would appear. This idea can also be developed as a standalone platform with a large database of examples, where an editor would be built alongside the search functionality.

Multimodal search-by-sketch: Building machine learning models and algorithms that allow learners to query using a rough sketch or an outline of the mental model. This could ease the learning curve for novice learners who lack the vocabulary to query the chart by name but can articulate visually what they want to build.

4.2 Pedagogical Relevance Modeling

Defining relevance in an educational context requires moving beyond standard visualization retrieval metrics. Previous user studies indicate that, in addition to high-similarity matches, participants also prioritize the diversity of retrieved results [21]. This need for diverse examples is particularly heightened in learning environments. To support this, we need to develop specialized metadata schemas and thesauri that extend current indexing methods and similarity criteria [20] and capture pedagogical utility.

For example, a schema for pedagogical purpose could encode dimensions of critique and reasoning such as *Cognitive Principles* (e.g., Gestalt proximity) or *Common Pitfalls* (e.g., truncated axes). At the same time, a visualization thesaurus is crucial to bridge the vocabulary gap for learners with different backgrounds and different levels of visualization literacy, by mapping lay descriptions to formal design terminology. For instance, linking a query for "Manhattan plot" to "GWAS significance scatterplot," a specialized scatterplot used in genome-wide association studies (GWAS) to visualize the relationship between genomic variants and a specific trait. Establishing this semantic infrastructure is a prerequisite for retrieval systems that can surface not only relevant but also conceptually aligned visualizations.

4.3 Collaborative Educational Corpora

Finally, there is an opportunity to build shared, open-access corpora specifically for education. These repositories would link visualizations, datasets, code, and teaching notes, curated by the community to serve as the backend for next-generation educational retrieval systems. Closely related to modeling pedagogical relevance above, capturing these nuanced attributes often requires human expertise that automated extraction tools currently lack. A collaborative corpus would allow educators to contribute these annotations, flagging visualizations that have proven effective in the classroom or that illustrate specific misconceptions. This direction primarily strengthens the management stage of the data lifecycle, while indirectly scaffolding production and consumption by providing a sustainable, well-annotated pool of examples that learners can adapt and critique.

5 Conclusion

In this position paper, we have argued that visualization retrieval is a valuable resource for teaching and learning data literacy. Retrieval systems can transform static visualization collections into navigable design spaces that help learners connect informal, task-oriented intents to concrete visual and code implementations. We demonstrated that integrating retrieval into the data lifecycle empowers learners to move beyond passive consumption, enabling them to actively explore design trade-offs, expand their visual vocabulary,

and critique representations through comparison. Realizing this potential calls for rethinking how we design and evaluate visualization retrieval systems. Beyond conventional similarity measures, future work should prioritize pedagogical relevance, integrate retrieval more tightly with authoring environments, and invest in community-driven corpora that link visualizations to datasets, code, and teaching goals. Positioning visualization retrieval in this way opens up opportunities to better support inquiry-driven learning, helping learners not only locate relevant visualizations but also articulate their intentions and reason more effectively with data.

Acknowledgments

The authors acknowledge funding by the National Institutes of Health (R01HG011773).

References

[1] Ali Baigelenov, Prakash Shukla, and Paul Parsons. 2025. How Visualization Designers Perceive and Use Inspiration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Hannah K. Bako, Xinyi Liu, Leilani Battle, and Zhicheng Liu. 2023. Understanding how Designers Find and Use Data Visualization Examples. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 1048–1058. doi:10.1109/TVCG.2022.3209490

[3] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. 2018. Beagle: Automated Extraction and Interpretation of Visualizations from the Web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3173574.3174168

[4] Katy Börner, Andreas Bueckle, and Michael Ginda. 2019. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1857–1864.

[5] Jian Chen, Meng Ling, Rui Li, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Torsten Möller, Robert S Laramee, Han-Wei Shen, Katharina Wünsche, et al. 2021. Vis30k: A collection of figures and tables from ieee visualization conference publications. *IEEE Transactions on Visualization and Computer Graphics* 27, 9 (2021), 3826–3833.

[6] Qing Chen, Ying Chen, Ruishi Zou, Wei Shuai, Yi Guo, Jiazhe Wang, and Nan Cao. 2025. Chart2Vec: A Universal Embedding of Context-Aware Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 31, 4 (2025), 2167–2181. doi:10.1109/TVCG.2024.3383089

[7] Dazhen Deng, Yihong Wu, Xinhuan Shu, Jiang Wu, Siwei Fu, Weiwei Cui, and Yingcai Wu. 2023. VisImages: A Fine-Grained Expert-Annotated Visualization Dataset. *IEEE Transactions on Visualization and Computer Graphics* 29, 7 (2023), 3298–3311. doi:10.1109/TVCG.2022.3155440

[8] Elif E Firat, Alark Joshi, and Robert S Laramee. 2022. Interactive visualization literacy: The state-of-the-art. *Information Visualization* 21, 3 (2022), 285–310.

[9] Yan Holtz. 2024. The Python Graph Gallery. https://python-graph-gallery.com/. Accessed: 2026-02-14.

[10] Yan Holtz. 2024. The R Graph Gallery. https://r-graph-gallery.com/. Accessed: 2026-02-14.

[11] Enamul Hoque and Maneesh Agrawala. 2020. Searching the Visual Style and Structure of D3 Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1236–1245. doi:10.1109/TVCG.2019.2934431

[12] Kristin Hunter-Thomson. 2025. How Can We Help Students Explore Data in Their Sensemaking?(Data Literacy 101). *Science Scope* 48, 1 (2025), 7–11.

[13] Maeve Hutchinson, Radu Jianu, Aidan Slingsby, Jo Wood, and Pranava Madhyastha. 2025. Capturing Visualization Design Rationale. In *2025 IEEE Visualization and Visual Analytics (VIS)*. 231–235. doi:10.1109/VIS60296.2025.00052

[14] Hammad R Khan, Jeonghyun Kim, and Hsia-Ching Chang. 2018. Toward an understanding of data literacy. *iConference 2018 Proceedings* (2018).

[15] Jens Koenen, Marvin Petersen, Christoph Garth, and Tim Gerrits. 2024. DaVE - A Curated Database of Visualization Examples. In *2024 IEEE Visualization and Visual Analytics (VIS)*. 11–15. doi:10.1109/VIS55277.2024.00010

[16] Ard W Lazonder and Ruth Harmsen. 2016. Meta-analysis of inquiry-based learning: Effects of guidance. *Review of educational research* 86, 3 (2016), 681–718.

[17] Haotian Li, Yong Wang, Aoyu Wu, Huan Wei, and Huamin Qu. 2022. Structure-aware Visualization Retrieval. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA). ACM, New York, Article 409, 14 pages. doi:10.1145/3491102.3502048

[18] Mahbubul Majumder, Becky Brusky, Michelle Friend, Julie Dierberger, Sarah Moulton, Andrew W Swift, and Betty Love. 2025. Developing a Data Literacy and Visualization Service Learning Course that Fulfills Undergraduate Quantitative Literacy Requirements. *Journal of Statistics and Data Science Education* just-accepted (2025), 1–12.

[19] Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

[20] Huyen N. Nguyen and Nils Gehlenborg. 2025. Safire: Similarity Framework for Visualization Retrieval. In *2025 IEEE Visualization and Visual Analytics (VIS)*. 246–250. doi:10.1109/VIS60296.2025.00055

[21] Huyen N Nguyen, Sehi L'Yi, Thomas C Smits, Shanghua Gao, Marinka Zitnik, and Nils Gehlenborg. 2025. Geranium: Multimodal Retrieval of Genomics Data Visualizations. doi:10.31219/osf.io/zatw9_v6

[22] Michael Oppermann, Robert Kincaid, and Tamara Munzner. 2021. VizCommender: Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 495–505. doi:10.1109/TVCG.2020.3030387

[23] Jonathan C Roberts, Peter Butcher, and Panagiotis D Ritsos. 2025. From Data to Insight: Using Contextual Scenarios to Teach Critical Thinking in Data Visualisation. In *2025 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVIS)*. IEEE, 65–70.

[24] Vidya Setlur, Andriy Kanyuka, and Arjun Srinivasan. 2023. Olio: A Semantic Search Interface for Data Repositories. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. ACM, New York, Article 95, 16 pages. doi:10.1145/3586183.3606806

[25] Astrid van den Brandt, Sehi L'Yi, Huyen N. Nguyen, Anna Vilanova, and Nils Gehlenborg. 2025. Understanding Visualization Authoring Techniques for Genomics Data in the Context of Personas and Tasks. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1180–1190. doi:10.1109/TVCG.2024.3456298

[26] Skylar Sargent Walters, Arthea Valderrama, Thomas C. Smits, David Kouřil, Huyen N. Nguyen, Sehi L'Yi, Devin Lange, and Nils Gehlenborg. 2025. GQVis: A Dataset of Genomics Data Questions and Visualizations for Generative AI. (Jul 2025).

[27] Shishi Xiao, Yihan Hou, Cheng Jin, and Wei Zeng. 2023. WYTIWYR: A User Intent-Aware Framework with Multi-modal Inputs for Visualization Retrieval. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 311–322. doi:10.1111/cgf.14832

[28] Yilin Ye, Rong Huang, and Wei Zeng. 2024. VISAtlas: An Image-Based Exploration and Query System for Large Visualization Collections via Neural Image Embedding. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (2024), 3224–3240. doi:10.1109/TVCG.2022.3229023

[29] Lu Ying, Aoyu Wu, Haotian Li, Zikun Deng, Ji Lan, Jiang Wu, Yong Wang, Huamin Qu, Dazhen Deng, and Yingcai Wu. 2024. VAID: Indexing View Designs in Visual Analytics System. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. ACM, New York, Article 198, 15 pages. doi:10.1145/3613904.3642237