Standard Practices for Data Processing and Multimodal Feature Extraction in Recommendation with DataRec and Ducho (D&D4Rec)

Alberto Carlo Maria Mancino
Politecnico di Bari
Bari, Italy
alberto.mancino@poliba.it

Matteo Attimonelli
Politecnico di Bari
Bari, Italy
Sapienza Università di Roma
Roma, Italy
matteo.attimonelli@poliba.it

Angela Di Fazio
Politecnico di Bari
Bari, Italy
angela.difazio@poliba.it

Daniele Malitesta
CentraleSupélec, Inria, Université Paris-Saclay
Gif-sur-Yvette, France
daniele.malitesta@centralesupelec.fr

Tommaso Di Noia
Politecnico di Bari
Bari, Italy
tommaso.dinoia@poliba.it

Abstract

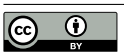
Recommendation pipelines involve several stages that can critically affect performance and reproducibility. However, early pipeline stages remain under-standardized, limiting comparability and interoperability across studies. This tutorial addresses this gap by providing both theoretical insights and hands-on experience with tools and practices for standardized data processing in recommender systems. In the first part, we introduce DataRec, a Python library for reproducible and interoperable data management, and discuss data filtering, splitting, and topological analysis techniques. In the second part, we explore multimodal feature extraction in domains such as fashion, music, and movies, focusing on the challenges of meaningful multimodal integration. We introduce Ducho, a unified framework for extracting audio, visual, and textual features using modern backends, and demonstrate its integration with the evaluation framework Elliot. The tutorial targets researchers and practitioners with an interest in recommender systems, data preprocessing, and multimodal modeling. All materials, including slides, code, datasets, and recordings, will be openly available on a dedicated tutorial website: https://sites.google.com/view/dd4rec-tutorial/.

CCS Concepts

• **Information systems** → **Personalization**; **Recommender systems**; *Extraction, transformation and loading*; *Multimedia and multimodal retrieval*; • **Software and its engineering** → *Software libraries and repositories.*

Keywords

Recommendation Datasets, Multimodal Recommendation, Reproducibility, Python Library



RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748009

ACM Reference Format:
Alberto Carlo Maria Mancino, Matteo Attimonelli, Angela Di Fazio, Daniele Malitesta, and Tommaso Di Noia. 2025. Standard Practices for Data Processing and Multimodal Feature Extraction in Recommendation with DataRec and Ducho (D&D4Rec). In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3705328.3748009

1 Motivation and Scope

The recommendation pipeline spans from dataset selection to model evaluation. While performance often depends on model design, other stages, like preprocessing and evaluation protocols, can significantly impact results, affecting comparability and reproducibility [4, 11, 13]. In this respect, notable effort has been devoted to model reproducibility [8, 15] and fair evaluation practices [7, 10, 22, 23]. Conversely, less attention has been focused on the initial stages of the recommendation pipeline, particularly dataset selection and **data processing**, which still lack unified and **standardized practices**. Indeed, such standardization could represent the cornerstone for interoperability and accelerating research progress in recommendation. To this end, our proposed tutorial "**Standard Practices for Data Processing and Multimodal Feature Extraction in Recommendation with DataRec and Ducho**" dubbed as **D&D4Rec** aims to fill in this literature gap, by providing a theoretical and practical overview of standard practices for data processing in recommendation, by ultimately discussing extraction procedures for **multimodal features** in specific tasks and scenarios.

In the first part of the tutorial, we will discuss the importance of standardization in recommendation pipelines, emphasizing the main limitations of current practices. We will then provide an overview of widely used recommendation datasets and common filtering and splitting strategies across various domains and recommendation paradigms [21]. We will also cover the structural characterization of recommendation datasets, especially topological properties of the user-item graph data, which have been shown to impact recommender system performance significantly [1, 9, 19].

In this respective, the Python library **DataRec** [21] has recently been introduced. It serves as a shared foundation for reproducible and interoperable data management in recommender systems. Unlike other established frameworks, DataRec is designed to be easily integrable into standalone projects, offering interfaces for exporting datasets in formats compatible with existing tools. Thus, the first hands-on session of our tutorial will illustrate how to effectively use DataRec and integrate it into any recommendation pipeline.

In the second part of the tutorial, we focus on multimodal feature extraction for recommendation. In domains like fashion, music, and movies, the multi-faceted features characterizing products and services may influence each customer on online selling platforms differently, paving the way to novel multimodal recommendation models that enhance the traditional recommendation data through high-level features extracted from items' multimodal content [17]. Indeed, the lack of standardized procedures to extract and process multimodal features in recommendation may undermine their quality and meaningfulness. Following this, we will provide an overview of widely used multimodal recommendation datasets, extraction, processing, and fusion procedures [5].

Then, we will present **Ducho** [6, 18], a unified framework to streamline multimodal feature extraction in recommender systems. It integrates popular backends (e.g., TensorFlow, PyTorch, and Transformers), providing a shared interface for extracting and processing audio, visual, and textual features in a highly-customizable way. Conclusively, the second hands-on session will show how to run complete multimodal recommendation pipelines by jointly leveraging Ducho and Elliot [3], a popular framework for recommender systems reproducibility and evaluation.

2 Tutorial Outline

Duration: **180 minutes (3 hours)**.

- **Introduction** → *15 minutes*
  - **[Theory]** The need for standard procedures to handle recommendation datasets
  - Outline and scopes of the tutorial
- **Part 1: Standard practices for data processing in recommendation with DataRec** → *90 minutes*
  - **[Theory]** Overview of data handling and processing → *25 minutes*
  - **[Theory]** Understanding the characteristics of recommendation datasets → *15 minutes*
  - **[Theory]** DataRec: a Python library for standardized data management → *10 minutes*
  - **[Hands-on]** How to use DataRec to process your dataset → *40 minutes*
- **Break and Q&A** → *10 minutes*
- **Part 2: Standard practices for multimodal features extraction in recommendation with Ducho** → *55 minutes*
  - **[Theory]** Overview of multimodal features extraction in recommendation → *15 minutes*
  - **[Theory]** Ducho: a unified framework for the extraction of multimodal features → *10 minutes*
  - **[Hands-on]** How to use Ducho to process your multimodal dataset → *30 minutes*
- **Q&A and closing remarks** → *10 minutes*

3 Targeted Audience and Useful Resources

This tutorial covers intermediate and advanced theoretical and practical topics, including recommendation dataset processing, characterization, and multimodal feature extraction for recommendation systems. It targets researchers and practitioners interested in these areas, even partially. While familiarity with Python and PyTorch is beneficial, step-by-step guidance will be provided, especially during hands-on sessions. The tutorial aims to equip attendees with rigorous and standardized theoretical and practical skills for managing multimodal recommendation datasets. Participants will receive access to presentation slides (via Slideshare), code and datasets for hands-on sessions (via GitHub and Google Colab), and a video recording of the tutorial (after the conference).

4 Main Differences with Previous Tutorials

We examined related tutorials presented at RecSys, UMAP, SIGIR, CIKM, WSDM, KDD, The Web Conference, and ECIR from 2021 to 2025 (when available), focusing on topics such as data processing, multimodal feature extraction, and standardization practices in recommendation. While recent efforts have addressed reproducibility in recommender systems, for instance through tutorials on offline evaluation protocols or practical deployment strategies (RecSys 2022 [7], RecSys 2023 [12], ECIR 2025 [14]), none have comprehensively tackled both standardized data handling and multimodal feature extraction within a unified, reproducible framework. Some tutorials, such as those presented at UMAP 2022 [16] and RecSys 2022 [2], have discussed standardization within framework-specific pipelines. These include systems designed to manage the full recommendation workflow—from content representation to evaluation or deployment—but they are closely tied to specific tools like ClayRS or industrial platforms like NVIDIA Merlin. In contrast, our tutorial introduces modular, lightweight libraries, DataRec and Ducho, designed to be easily integrated into any recommender pipeline, fostering openness, reproducibility, and interoperability across different research and development environments. Moreover, we have previously covered the topological characterization of recommendation datasets in a tutorial at LoG 2023 [20]. This new tutorial extends that work by addressing standardization and multimodality more broadly, across different tasks and data modalities.

Acknowledgments

This work has been carried out while *Matteo Attimonelli* was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with *Politecnico Di Bari*. This work was partially supported by the following projects: LUTECH DIGITALE 4.0. "VAI2C"- Virtual Artificial Intelligence Contact Center. "IDENTITA" CUP D93C22001020008. "REACH-XY": RESEARCH ACTIONS FOR REDUCING THE IMPACT ON AGRICULTURAL AND NATURAL ECOSYSTEMS OF THE HARMFUL PLANT PATHOGEN XYLELLA FASTIDIOSA, CUP B93C22001920001. "Patto Territoriale sistema universitario pugliese" CUP F61B23000370006- cod. id. PATTI_TERRITORIALI_WP1. Natuzzi S.p.A. del Contratto di Sviluppo Industriale ai sensi dell'art. 9 del Decreto del Ministro dello Sviluppo Economico del 09.12.2014.

References

[1] Gediminas Adomavicius and Jingjing Zhang. 2012. Impact of data characteristics on recommender systems performance. *ACM Trans. Manag. Inf. Syst.* 3, 1 (2012), 3:1–3:17.

[2] Ronay Ak, Benedikt Schifferer, Sara Rabhi, and Gabriel de Souza Pereira Moreira. 2022. Training and Deploying Multi-Stage Recommender Systems. In *RecSys*. ACM, 706–707.

[3] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR*. ACM, 2405–2414.

[4] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM*. ACM, 601–610.

[5] Matteo Attimonelli, Danilo Danese, Angela Di Fazio, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. 2024. Ducho meets Elliot: Large-scale Benchmarks for Multimodal Recommendation. *CoRR* abs/2409.15857 (2024).

[6] Matteo Attimonelli, Danilo Danese, Daniele Malitesta, Claudio Pomo, Giuseppe Gassi, and Tommaso Di Noia. 2024. Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation. In *WWW (Companion Volume)*. ACM, 1075–1078.

[7] Francesco Barile, Amra Delic, and Ladislav Peska. 2022. Tutorial on Offline Evaluation for Group Recommender Systems. In *RecSys*. ACM, 702–705.

[8] Alejandro Bellogín and Alan Said. 2021. Improving accountability in recommender systems research through reproducibility. *User Model. User Adapt. Interact.* 31, 5 (2021), 941–977.

[9] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?. In *WSDM*. ACM, 141–149.

[10] Charles L. A. Clarke, Fernando Diaz, and Negar Arabzadeh. 2023. Preference-Based Offline Evaluation. In *WSDM*. ACM, 1248–1251.

[11] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *RecSys*. ACM, 101–109.

[12] Kim Falk and Morten Arngren. 2023. Recommenders In the wild - Practical Evaluation Methods. In *RecSys*. ACM, 1.

[13] Antonio Ferrara, Angela Di Fazio, Alberto Carlo Maria Mancino, Tommaso Di Noia, and Eugenio Di Sciascio. 2025. Enhancing Utility in Differentially Private Recommendation Data Release via Exponential Mechanism. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 15574)*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer, 34–51. doi:10.1007/978-3-031-88714-7_3

[14] Antonio Ferrara, Claudio Pomo, and Nicola Tonellotto. 2025. Enhancing Reproducibility and Replicability in Information Retrieval: A Path Towards Scientific Integrity and Effective Research. In *ECIR (5) (Lecture Notes in Computer Science, Vol. 15576)*. Springer, 266–272.

[15] Juliana Freire, Philippe Bonnet, and Dennis E. Shasha. 2012. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *SIGMOD Conference*. ACM, 593–596.

[16] Pasquale Lops, Cataldo Musto, and Marco Polignano. 2022. Semantics-aware Content Representations for Reproducible Recommender Systems (SCoRe). In *UMAP*. ACM, 354–356.

[17] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. 2025. Formalizing Multimedia Recommendation through Multimodal Deep Learning. *Trans. Recomm. Syst.* 3, 3 (2025), 37:1–37:33.

[18] Daniele Malitesta, Giuseppe Gassi, Claudio Pomo, and Tommaso Di Noia. 2023. Ducho: A Unified Framework for the Extraction of Multimodal Features in Recommendation. In *ACM Multimedia*. ACM, 9668–9671.

[19] Daniele Malitesta, Claudio Pomo, Vito Walter Anelli, Alberto Carlo Maria Mancino, Tommaso Di Noia, and Eugenio Di Sciascio. 2024. A Novel Evaluation Perspective on GNNs-based Recommender Systems through the Topology of the User-Item Graph. In *RecSys*. ACM, 549–559.

[20] Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. 2023. Graph Neural Networks for Recommendation: Reproducibility, Graph Topology, and Node Representation. *CoRR* abs/2310.11270 (2023).

[21] Alberto Carlo Maria Mancino, Salvatore Bufi, Angela Di Fazio, Antonio Ferrara, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. 2025. DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender Systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3478–3487. doi:10.1145/3726302.3730320

[22] Aixin Sun. 2023. On Challenges of Evaluating Recommender Systems in an Offline Setting. In *RecSys*. ACM, 1284–1285.

[23] Eva Zangerle and Christine Bauer. 2023. Evaluating Recommender Systems: Survey and Framework. *ACM Comput. Surv.* 55, 8 (2023), 170:1–170:38.