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

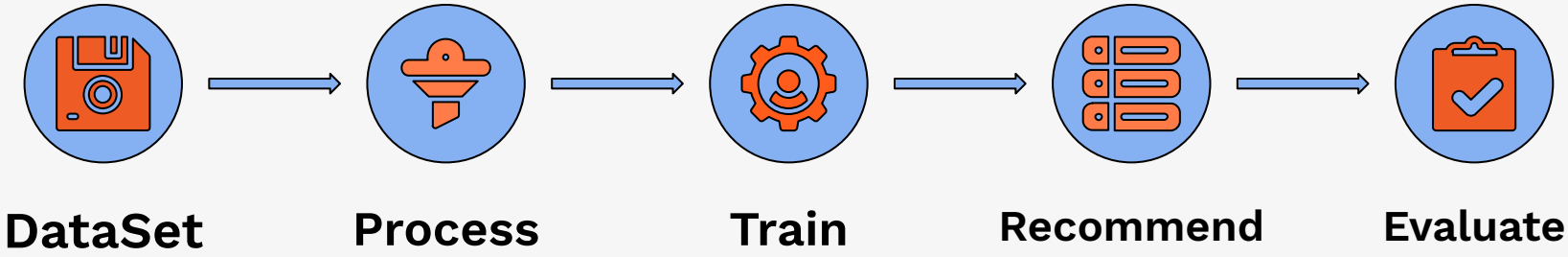
Politecnico di Bari

SAPIENZA
UNIVERSITÀ DI ROMA

CentraleSupélec

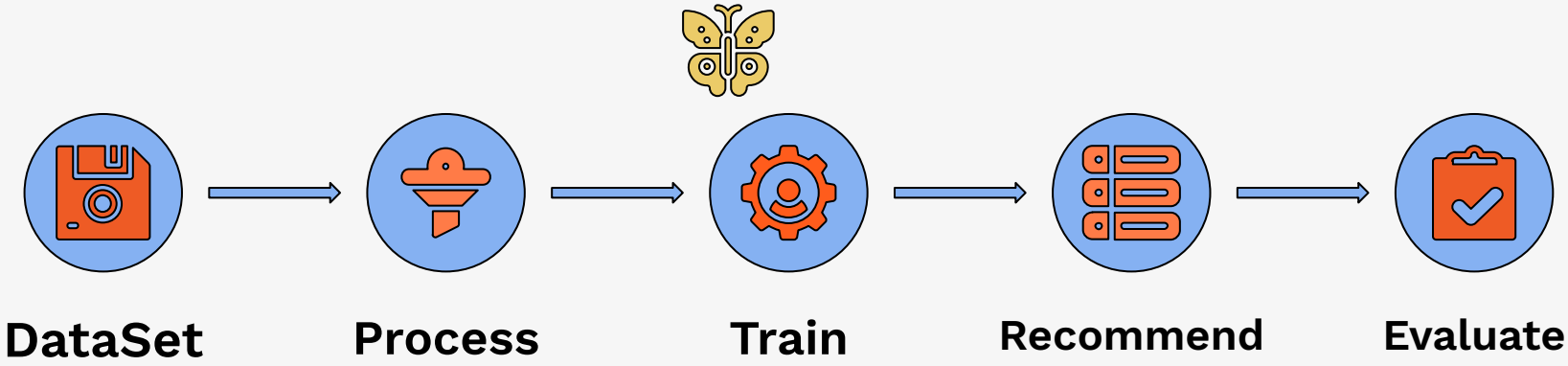
université PARIS-SACLAY

Offline Evaluation Pipeline



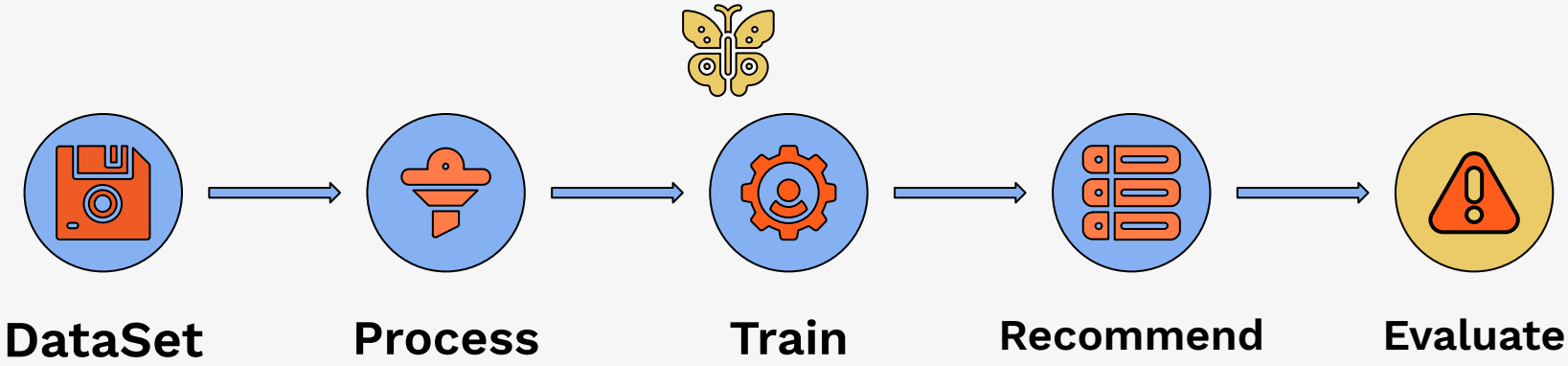
Typical pipeline for offline evaluation in recommender systems.

Offline Evaluation Pipeline



Without **standardization**, small changes may cause butterfly effect.

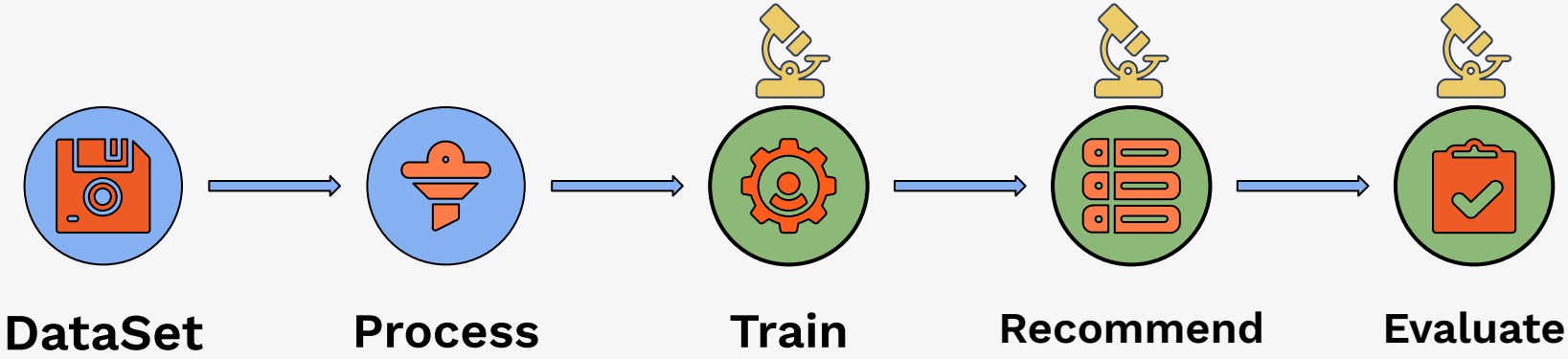
Offline Evaluation Pipeline



Small **variations** at any stage can affect the results, compromising **reproducibility** and **fair comparison**. [1,2]

[1] Armstrong et al., CIKM 2009: “Improvements that don't add up: ad-hoc retrieval results since 1998
[2] Ferrari Dacrema et al., RecSys 2019: “Are we really making much progress? A worrying analysis of recent neural recommendation ...”

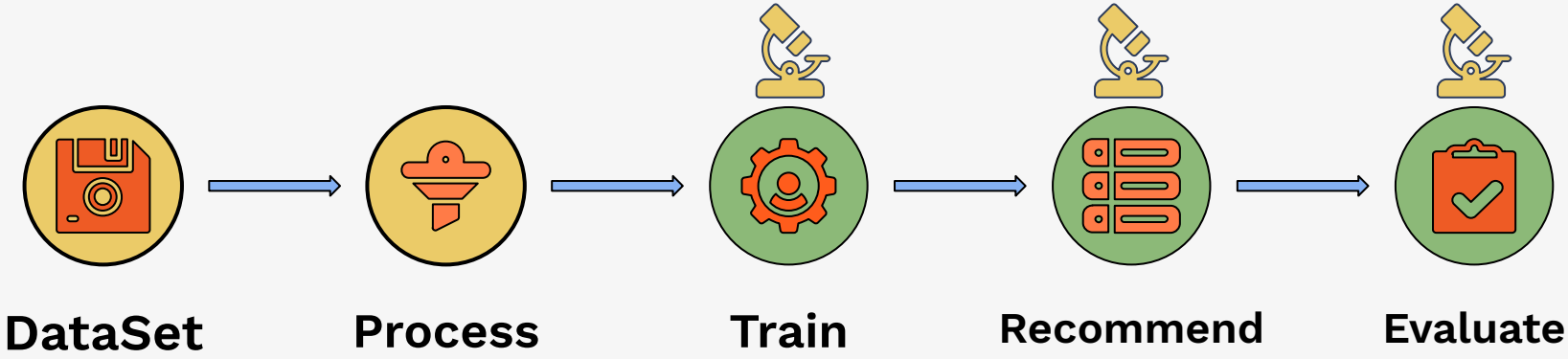
Offline Evaluation Pipeline



So far, notable efforts has been devoted to guarantee models **reproducibility** [3] and fair **evaluation practices** [4].

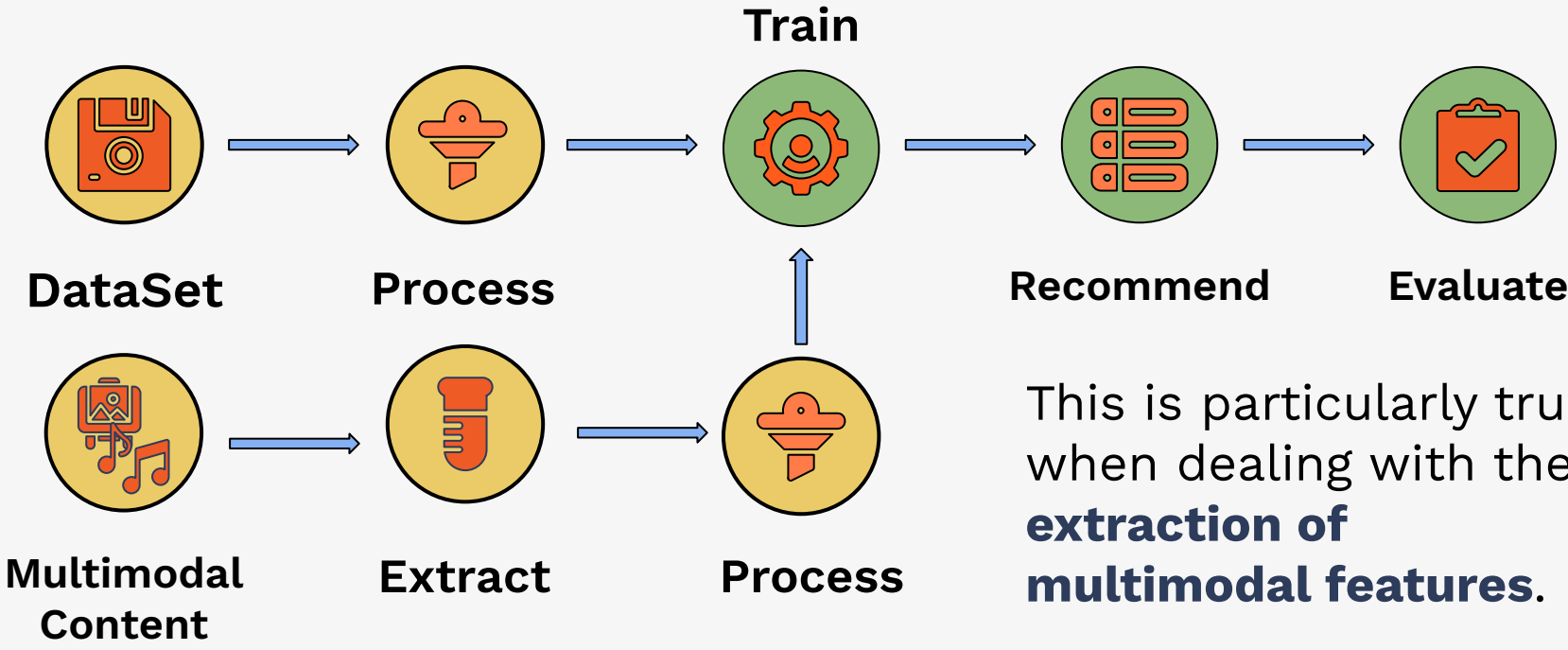
[3] Bellogín & Said, UMUAI 2021: "Improving accountability in recommender systems research through reproducibility"
[4] Sun, RecSys 2023: "On Challenges of Evaluating Recommender Systems in an Offline Setting"

Offline Evaluation Pipeline



Less attention has been devoted to the first stages of the pipeline, in particular the **data selection** and the **data processing**.

Offline Evaluation Pipeline



This is particularly true when dealing with the **extraction of multimodal features**.

D&D4Rec aims to fill this gap

D&D4Rec

With this tutorial we provide an overview of the ***most common practices*** *for data selection, filtering, processing and splitting in recommendation.*

Moreover, we present ***multimodal feature extraction*** *procedures for recommendation.*

D&D4Rec

*Not only theory but also practice with **two hands-on***

DataRec Hands-on

Exploring how to use and integrate the DataRec **Python library** for data selection and processing [5].

Ducho Hands-on

Presenting a **unified framework** designed to streamline multimodal feature extraction for recommenders [6].

[5] Mancino et al., SIGIR 2025: "DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender ..."
[6] Attimonelli et al., WWW 2024: "Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal ..."

&D4Rec

DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender Systems

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

Salvatore Bufi
s.bufi@phd.poliba.it
Politecnico di Bari
Bari, Italy

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

Antonio Ferrara
antonio.ferrara@poliba.it
Politecnico di Bari
Bari, Italy

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

Claudio Pomo
claudio.pomo@poliba.it
Politecnico di Bari
Bari, Italy

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

hands-on

nds-on

g a unified

k designed to streamline multimodal feature extraction for recommenders [6].

Python library for data selection and processing [5].

[5] Mancino et al., SIGIR 2025: "DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender ..."
[6] Attimonelli et al., WWW 2024: "Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal ..."

&D4Rec

hands-on

nds-on

Pyth

selec

DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender Systems

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

Salvatore Bufi
s.bufi@phd.poliba.it
Politecnico di Bari
Bari, Italy

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

Antonio Ferrara
antonio.ferrara@poliba.it
Politecnico di Bari
Bari, Italy

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

Claudio Pomo
claudio.pomo@poliba.it
Politecnico di Bari
Bari, Italy

Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation

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

Danilo Danese
Politecnico di Bari, Italy
danilo.danese@poliba.it

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

Claudio Pomo
Politecnico di Bari, Italy
claudio.pomo@poliba.it

Giuseppe Gassi
Politecnico di Bari, Italy
g.gassi@studenti.poliba.it

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

[5] Mancino et a
[6] Attimonelli et al.,

D&D4Rec

Standard Practices for Data Processing and Multimodal Feature Extraction in Recommender Systems Offline Evaluation

ALBERTO CARLO MARIA MANCINO, Politecnico di Bari, Italy
MATTEO ATTIMONELLI, Politecnico di Bari, Italy
ANGELA DI FAZIO, Politecnico di Bari, Italy
DANIELE MALITESTA, Université Paris-Saclay, CentraleSupélec, Inria, France
TOMMASO DI NOIA, Politecnico di Bari, Italy

recommenders [6].

[5] Mancino et al., SIGIR 2025: "DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender ..."
[6] Attimonelli et al., WWW 2024: "Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal ..."

Tutorial Overview

01 *Introduction*

02 *Data Handling and Processing*

03 *Data Characteristics*

04 *Introduction to DataRec*

05 ***Hands-on Session:*** *DataRec*

06 *Multimodal Feature Extraction*

Tutorial Overview

05 ***Hands-on Session:*** *DataRec*

06 *Multimodal Feature Extraction*

07 *Introduction to Ducho*

08 ***Hands-on Session:*** *Ducho*

09 *Final Remarks*

Presenters

Alberto C. M. Mancino
Politecnico di Bari

Matteo Attimonelli
Politecnico di Bari

Angela Di Fazio
Politecnico di Bari

Daniele Malitesta
Centrale Supélec

Tommaso Di Noia
Politecnico di Bari