

ArtiFact: A Large-Scale Multi-Modal Cultural Heritage Dataset

Luciano Duarte

Olga Ovcharenko

Sebastian Schelter

MOTIVATION Why Is Another Dataset Needed?

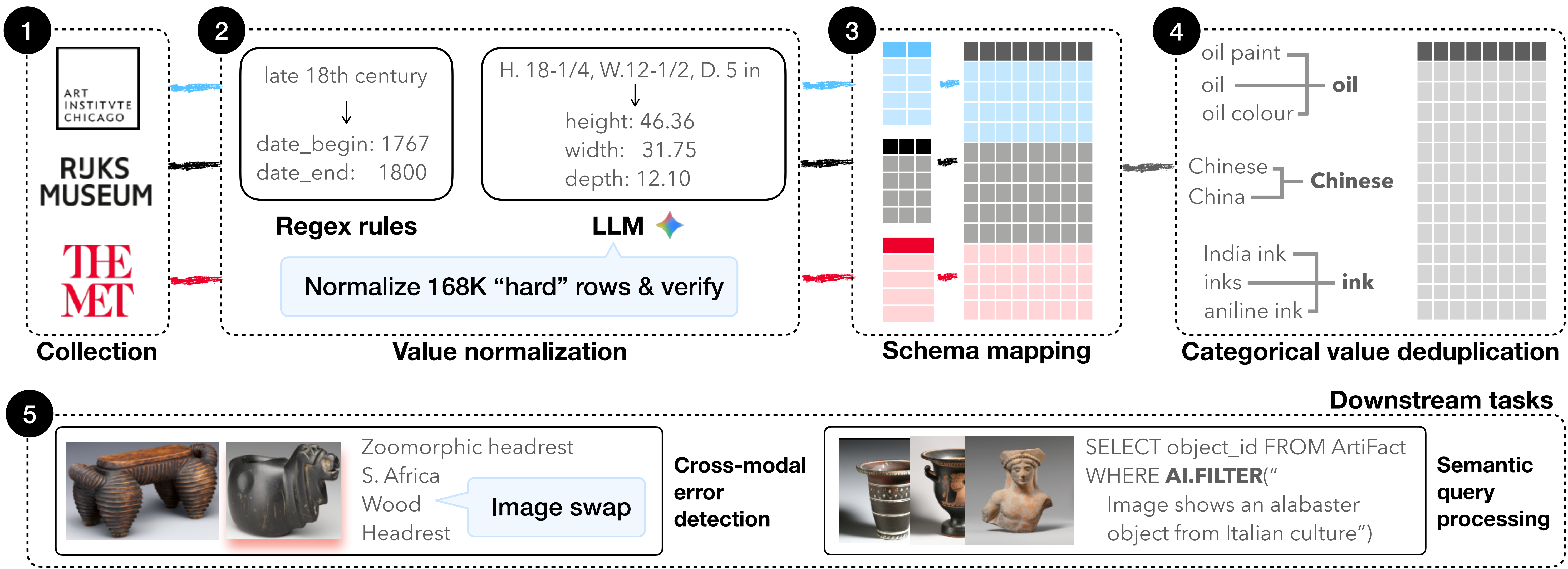
Multi-modal data management is a central problem, spanning data integration, semantic query processing, entity resolution, and data quality assessment. However, there is a **lack of well-curated large-scale multi-modal datasets**:

- No large-scale real-world dataset for cultural heritage.
- Multi-modal data is everywhere, but it is noisy.

DATASET How Was the Data Collected and Prepared?

ArtiFact is a large-scale multi-modal cultural heritage dataset of 651,045 museum objects with tables, images, and text.

- Open-access records** from the Metropolitan Museum of Art, the Art Institute of Chicago, and Rijksmuseum.
- A unified data preparation pipeline** to parse and normalize heterogeneous museum records.



A curated error taxonomy informed by museum curators for the **cross-modal error detection** downstream task finds errors in tabular data by jointly evaluating multiple modalities.

Thomas Jefferson
Silver (actual: **Copper**)
1801
J. M. Reich

1 Physical Error

Whistle
Kuba (actual: **Kongo**)
Wood, wire
DR Congo

2 Culture Error

Covered ovoid jar
1600-1624
(actual: **1700-1724**)
Porcelain, jade

3 Temporal Error

Seated musician
China
Tang Dynasty
14.9 cm
Earthenware

7 Image swap

The Actor Arashi III
Kunihisa (actual: **Yoshikatzu**)
Woodblock print

4 Identity Error

Baz
Iran (actual: **Egypt**)
Metal, skin, shell

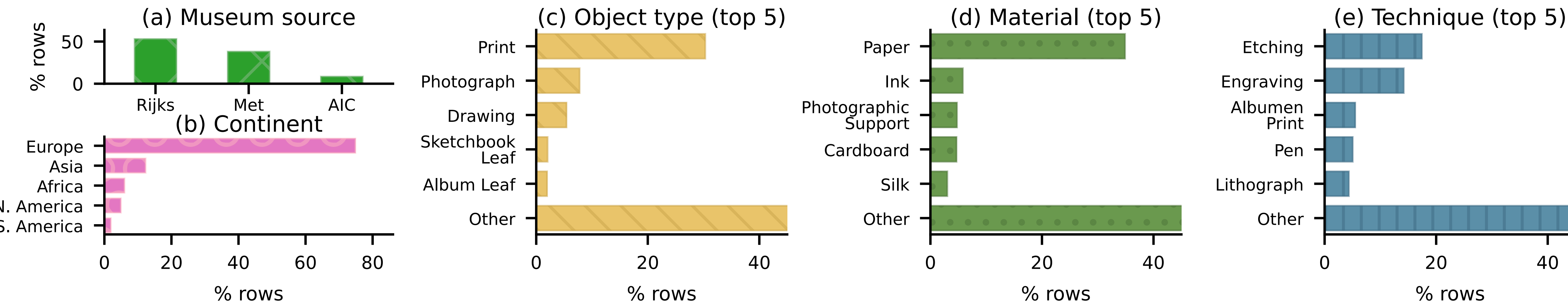
5 Geographic Error

Monkey with a young Toy
1755
57 x 95 cm
(actual: **5.7 x 9.5 cm**)
Soft paste porcelain
English
Chelsea Porcelain

6 Spatial Error

INSIGHTS What Can We Learn From the Dataset?

European objects dominate the dataset. Dutch/German/French prints on paper from 19th century are the most common artifacts.



LLMs have poor historical knowledge. Visual ambiguity and cultural nuances and biases cause errors in downstream tasks.



Contact

ArtiFact