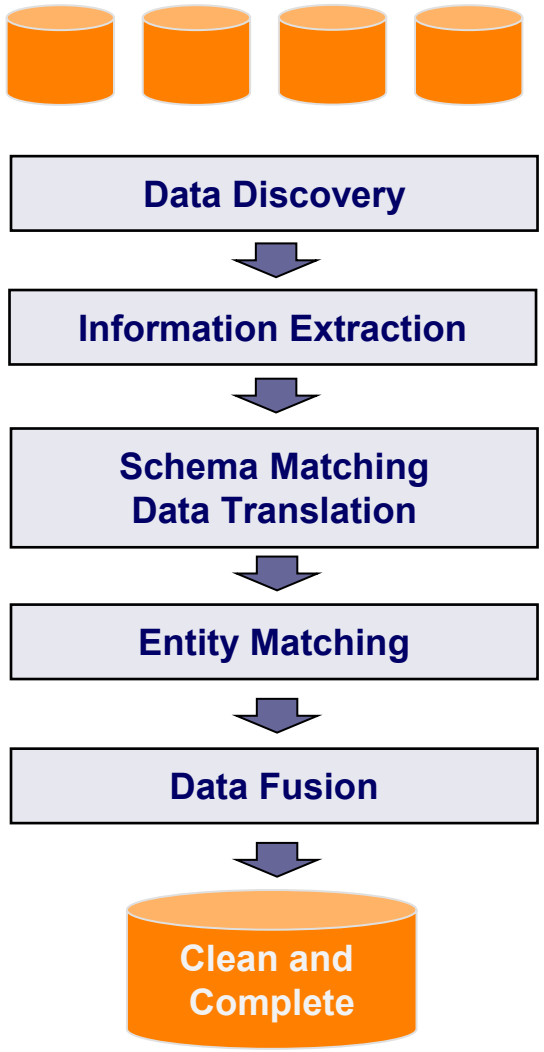
LLM-Agents in Data Integration Pipelines

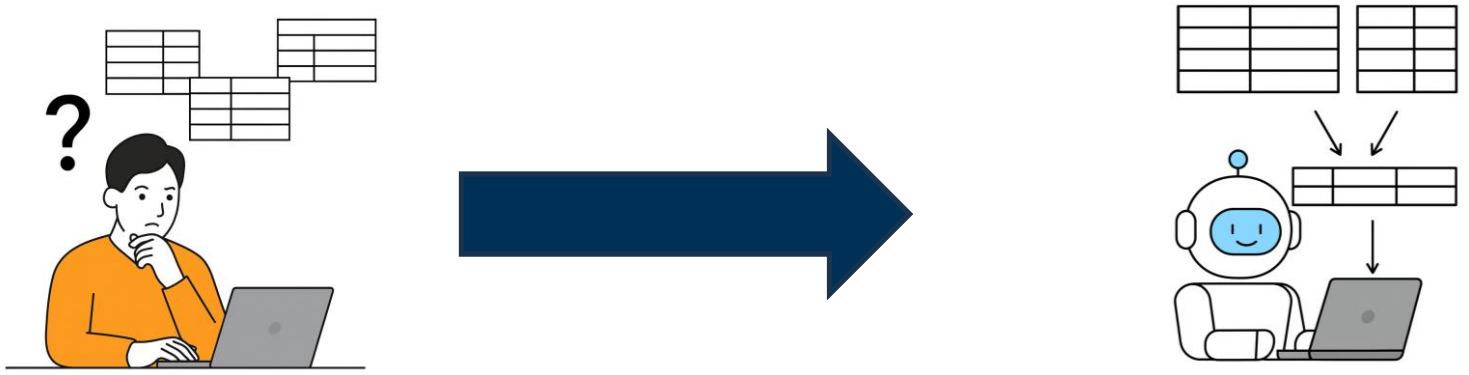
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



- Data Integration is a key step in most data science projects:
  - Corporate context: Mergers
  - Web context: Google News, Price comparison, Recommendation Systems
- Data integration pipelines involve multiple steps
- Errors accumulate over the steps and often only become clear during data fusion

From Manual Data Integration...
To Self-Improving Pipelines

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



- Data integration = iterative, messy, repetitive, error-prone
- Manual configuration of schema matching, entity resolution, …
- Manual error analysis

- LLM agents read execution reports and perform error analysis
- Agents improve code of integration pipeline to fix errors
- Using code synthesis (Embrace the vibes;))

This project explores how LLM agents can program, evaluate, and improve integration pipelines by themselves

Project Goal: Implement, evaluate and refine a data integration Agent

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Prof. Bizer

Aaron Steiner

Involves:

- Select data integration use cases that agents need to solve
- Implement and evaluate coding agents

Questions to answer:

- How good are agents at automatically performing error analysis?
- How good are agents at self-refining data integration workflows?

Requirements:

- Programming skills in Python
- Relevant courses: LLMs and Agents, Web Data Integration recommended

Organization: English, 5 people, 6 months