Miniaturizing Data Harmonization
Methods to Facilitate Training in the OMOP Data Ecosystem

PRESENTER: **Gesquiere** Emma

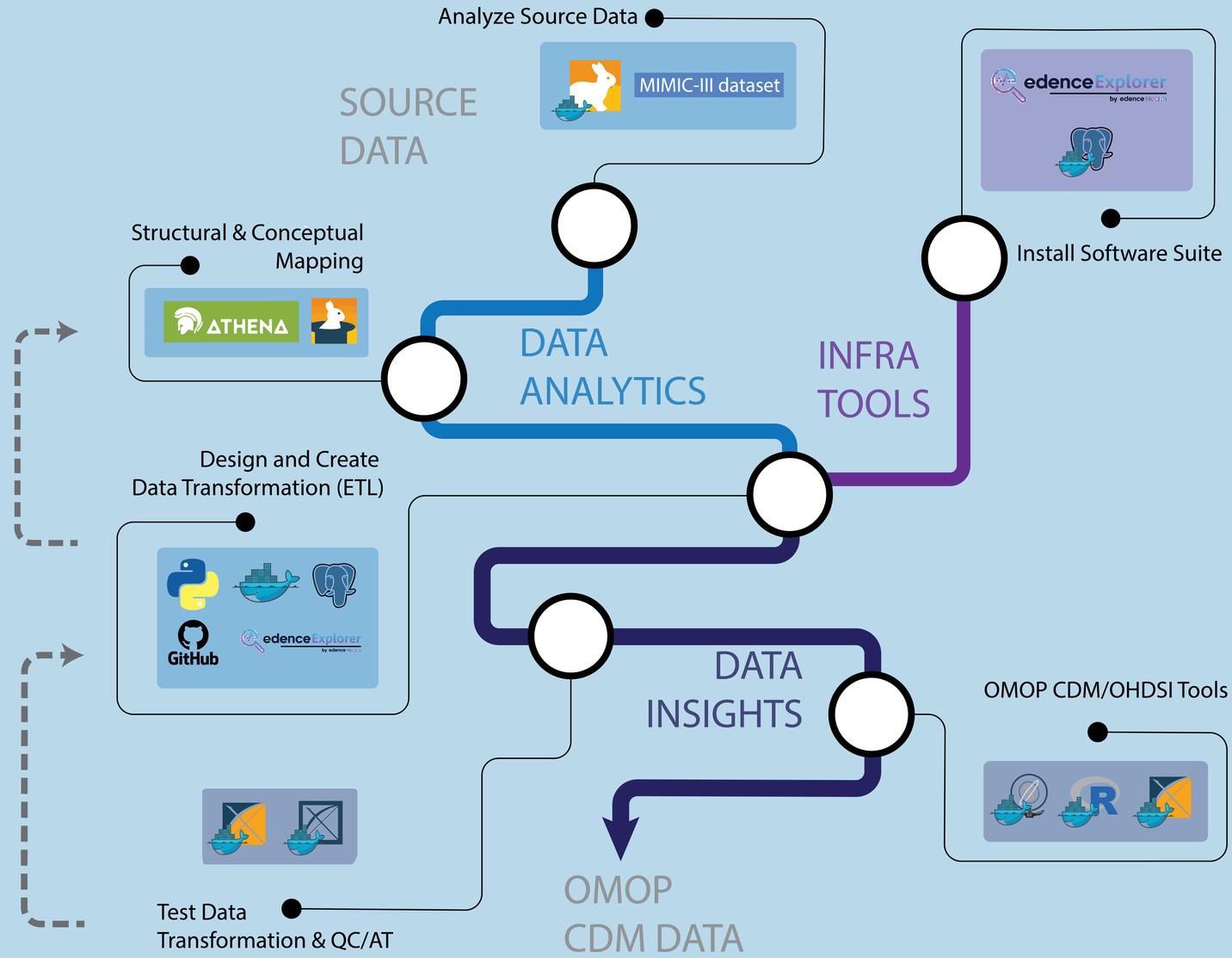
INTRO:

- **Who cares?**
  Currently available resources like EHDEN Academy and The Book of OHDSI provide a wealth of information on data harmonization and its caveats; these resources, however, often present different aspects of the process as independent and do not always convey practical interdependencies that are ultimately critical for successful harmonization.

METHODS

1. Trainee harmonizes a small set of synthetically generated medical data (MIMIC) into OMOP CDM format.
2. Utilizes repackaged open source OHDSI and EHDEN software/tools as Docker images to support the harmonization process.
3. Develops several transformations of an Extract-Transform-Load (ETL) process, which are Dockerized, orchestrated in Python 3, and are executed either as (1) Pandas dataframes, (2) database operations via direct SQL queries, or (3) embedded datatable objects within SQLalchemy.

We created a **miniature harmonization training** method with **dockerized software packages** to support the adoption of the OMOP CDM.



Scan the QR or go to edence.health/publications/ to learn more about the mini-harmonization project as well as other projects from edenceHealth

RESULTS

- We hope to achieve:
  - increased knowledge of containerization as a useful and efficient deployment modality for harmonization projects
  - increased accessibility and understanding of the OMOP CDM with regard to extracting insight from real-world evidence
  - increased number of proficient users of OHDSI and EHDEN software packages
- All components of this project are available free of charge and can be executed directly on the trainee's machine following a free installation of the Docker engine.
- We plan to extend this tool to the broader OHDSI community by sharing the pre-built Docker images and associated project documentation.

Currently repackaged tools include White Rabbit, Achilles, Data Quality Dashboard, Atlas and WebAPI, as well as network-independent Docker-images of R Studio and the Synthea medical data generation libraries.

Emma Gesquiere, Jared Houghtaling, Ivo Mbi Kubam, Frederic Jung, Ben Burke, Freija Descamps, Lars Halvorsen

edence Health

OHDSI