DEMOCRATIZING DATA & AI
FOR NHR

CARSTEN BINNIG

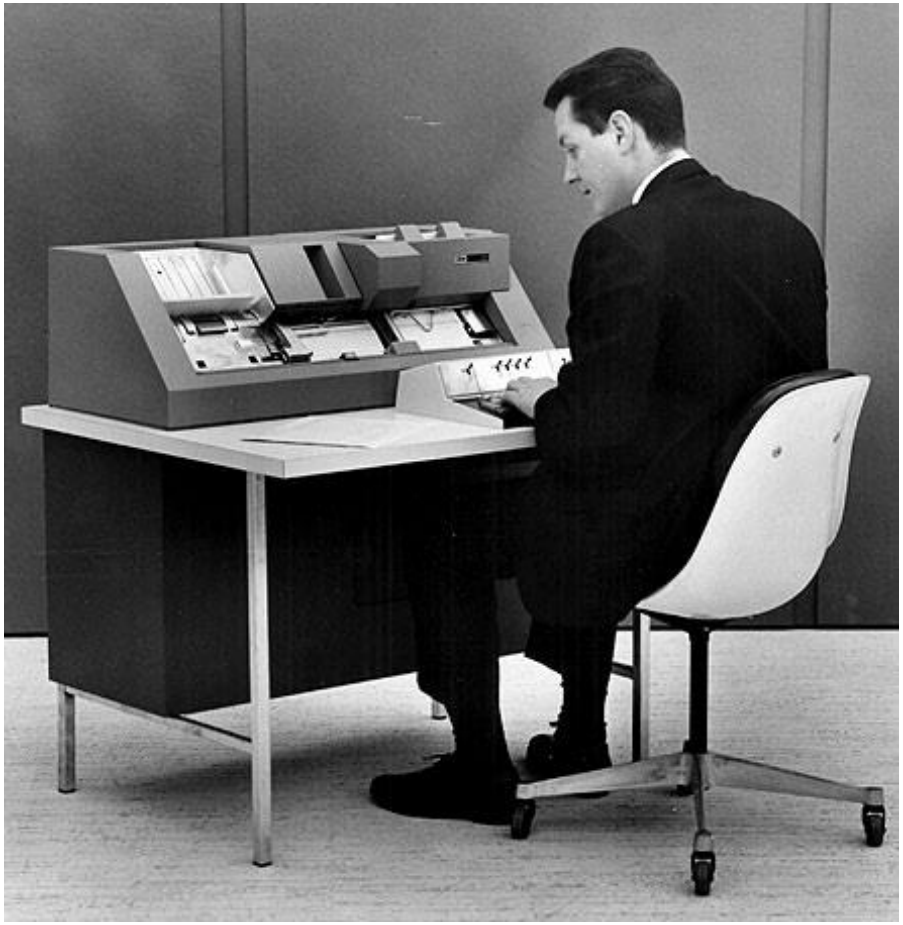
CSG DATA MANAGEMENT, NHR4CES, TU DARMSTADT

VISION: DEMOCRATIZING DATA & AI

TODAY'S END USER DEVICES

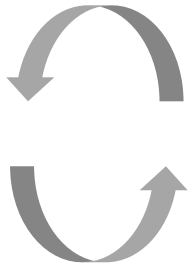
...

TODAY'S DATA & AI SYSTEMS!

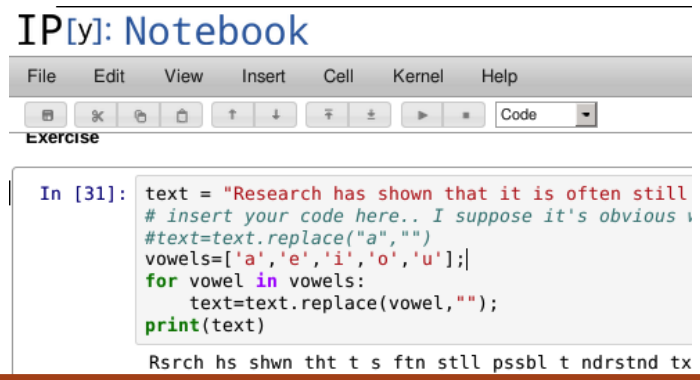


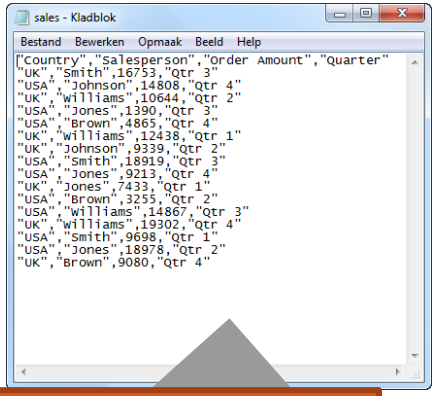
WHAT ARE THE PAIN POINTS?

Domain Expert



Data Scientist





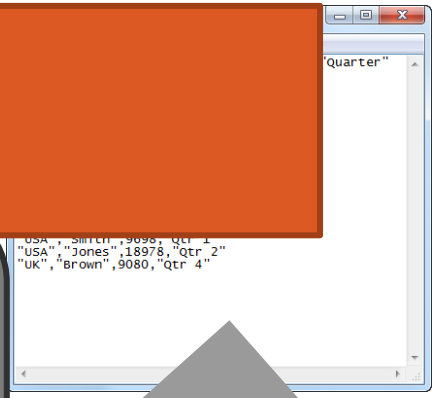
Research Agenda: Make Data & AI Systems **more useable, efficient, …. for NHR**

REVISIT DATA & AI SYSTEM STACK

Domain Expert

Data Scientist

2. User Efficiency
(Visual & NL interfaces)



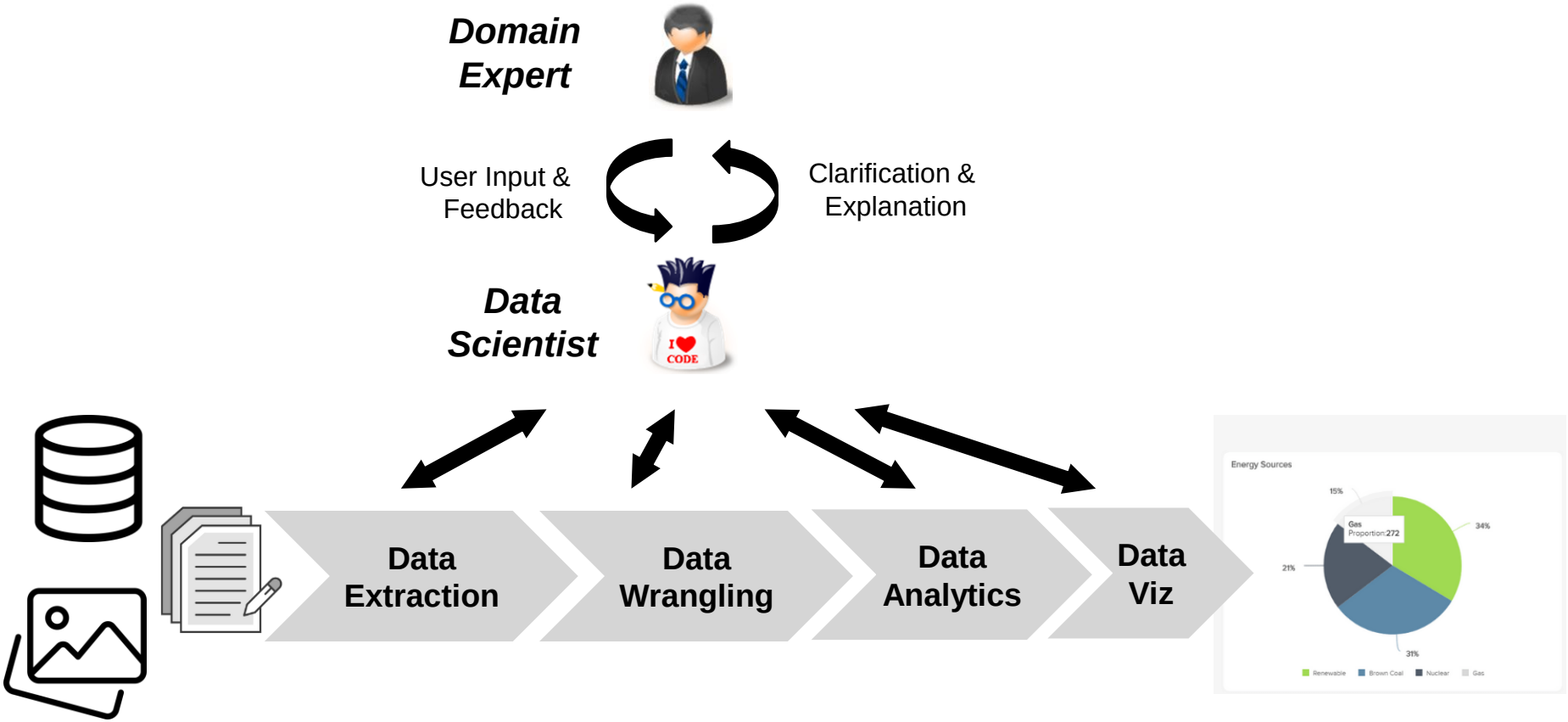
Focus: Automation of User Tasks (Data & AI Systems)

3. Automation
(System ... User Tasks)

1. System Efficiency
(Scalability on Modern Hardware)

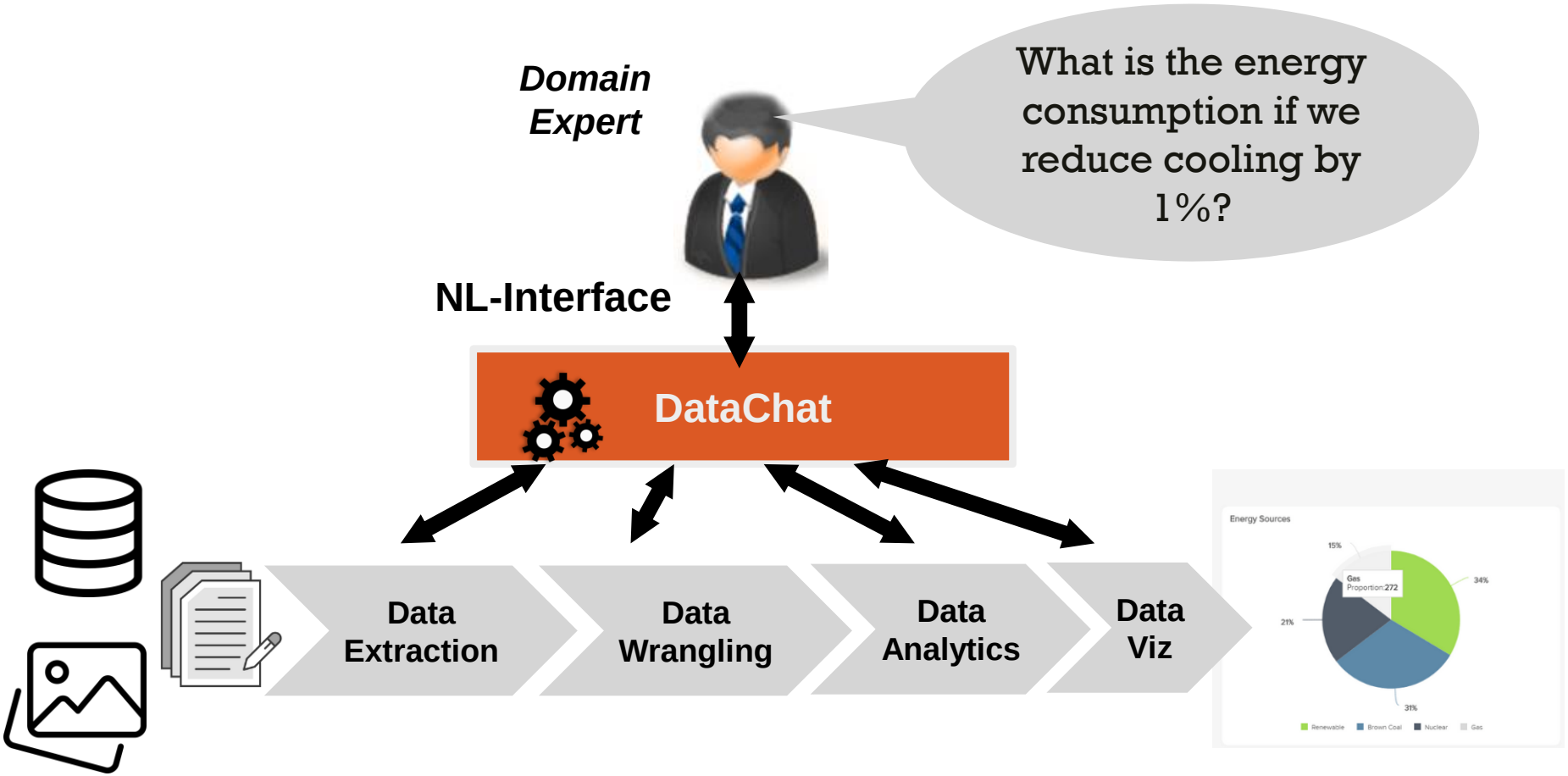
Manual Data Preparation

HOW ARE DATA & AI PIPELINES BUILT?



Today: Manually-composed "pipelines"
→ high overheads + expert skills needed

FUTURE: CONVERSATIONAL DATA & AI



Idea: Use AI to **automate construction pipelines**
→ reduce high manual overheads

DIRECTION 1: DATA EXTRACTION

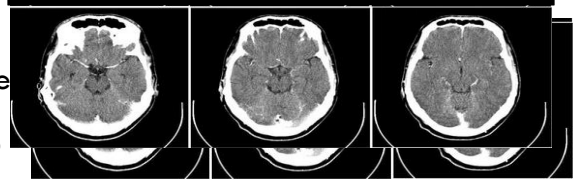
[SIGMOD'23, Another paper submitted]

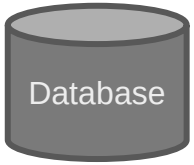
Patients (table)		
name	age	gender
Alice	42	f
Bob	23	m

Examinations (text)

Alice was diagnosed with fever …

Scans (images)





+

GPT-like Model

What is the age and diagnosis of patient Alice?

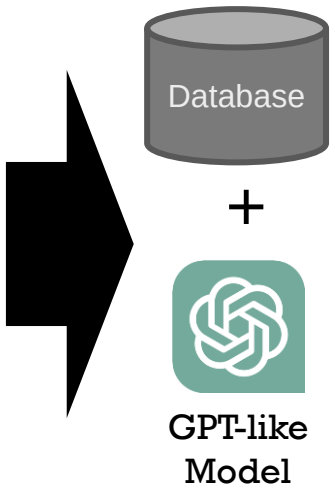
Use a GPT-like model to "**query**" a **mix of table, text, image data**

DIRECTION 2: DATA AUGMENATION

[AIDM@SIGMOD'23]

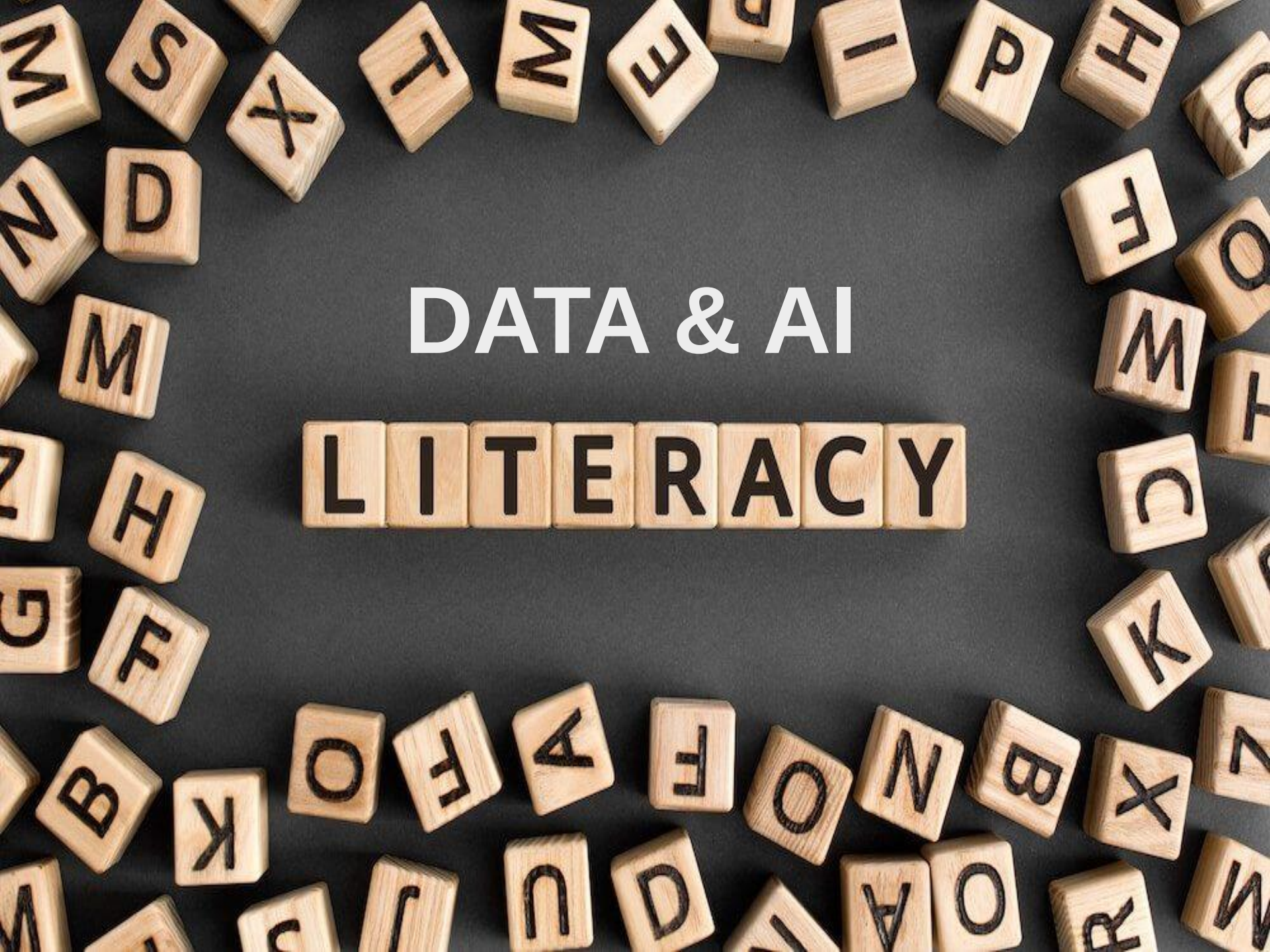
movies		
movie	**genre**	**revenue**
Avatar	Sci-Fi	2.74B $
Avatar2	Sci-Fi	2.24B $

actors (GPT-table / virtual)		
movie	**actor**	**age**
Avatar	Zoe S	…
Avatar	San W	…



What is the **average revenue** of movies with **actor „Zoe S“**?

Use a GPT-like model to **augment data with external "world-knowledge"**


DATA & AI
LITERACY

DATA & AI EDUCATION: HIGHLY INDIVIDUALIZED DEMANDS

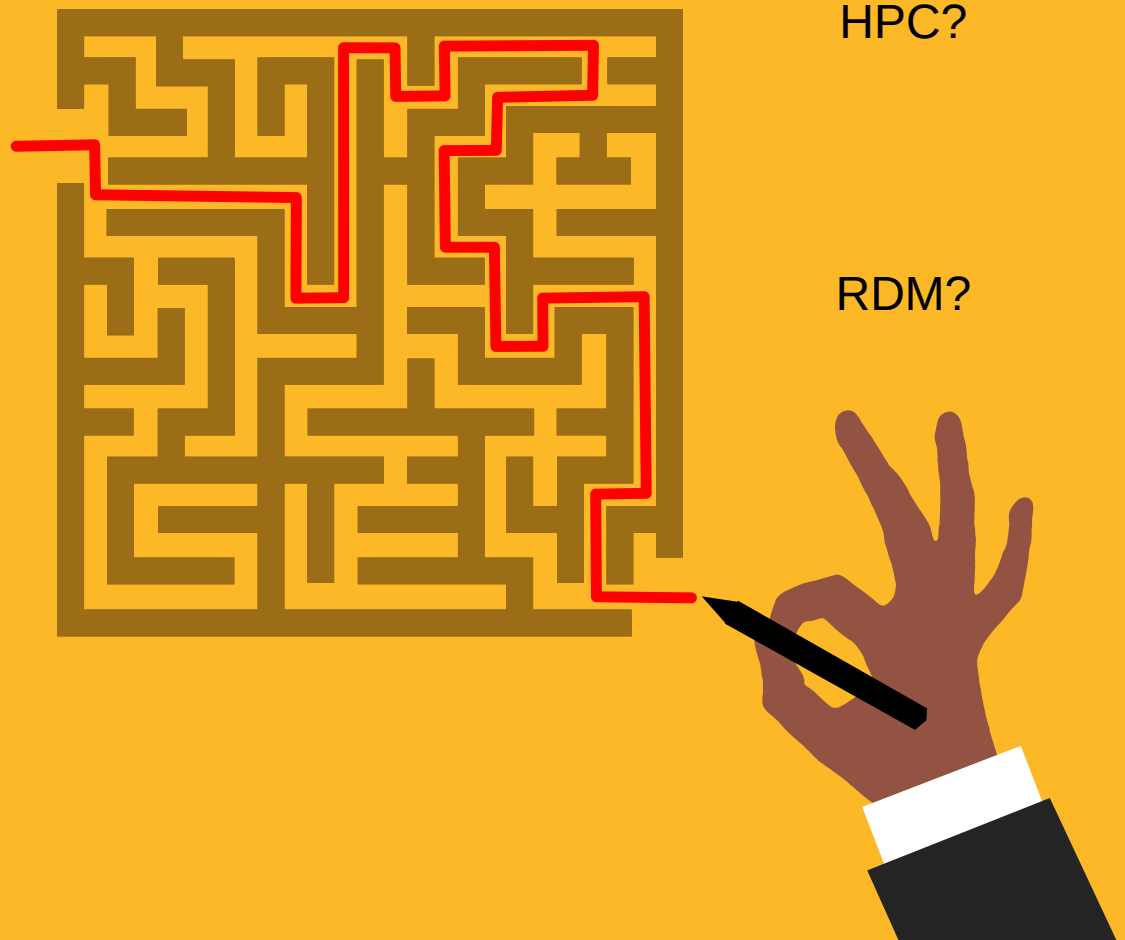
Python Programming?

HPC?

Data Engineering?

RDM?

ML/AI?



Where to
start?

DATA & AI LITERACY HUB 4 NHR

https://data-ai-literacy.ml/

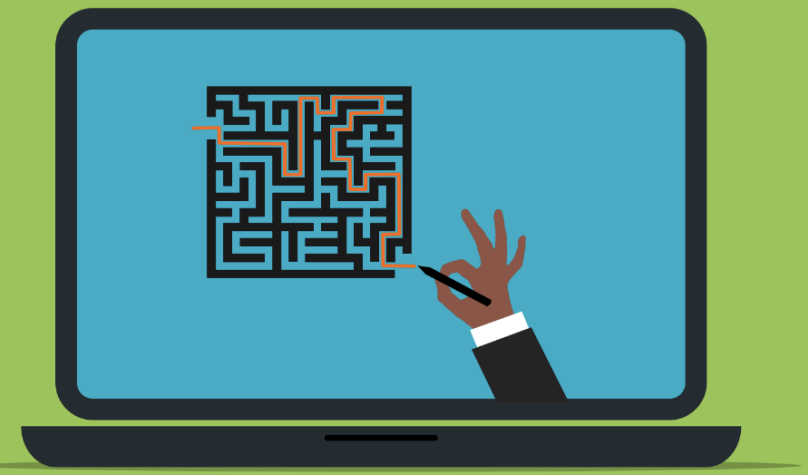
Start with one of our learning paths:

...

Not sure what to choose?

START LEARNING

The Data and AI Literacy Hub is a guided learning platform that offers individual learning paths for different backgrounds and skill levels. It enables you to develop the skills necessary to navigate the data-driven world and apply AI techniques in custom projects.



Learning Paths

Our carefully curated learning paths guide you in finding learning resources that fit your needs and level of knowledge.

Get started with HPC

Introduction to Programming

Learn the Basics of Data Science & AI

The essentials of RDM

Theoretical Foundations of AI

OPEN EDUCATION: EVERYONE CAN PROPOSE NEW PATHS & COURSES

```yaml
title: The Data Science Pipeline
slug: the_data_science_pipeline
source: Interactive Course from Prismia
description: |
  The beginner-friendly course „The Data Science Pipeline" is a great way to get an
   overview about all the essential steps in data science, from acquiring data to
   building and evaluation models. You'll get to know different sources to acquire data
   and learn ways to clean and wangle your data to make it ready for analysis. You'll get
    hands-on experience in building a first model to make predictions with popular data
   science tools and libraries.
targetgroups:
  - "Domain experts with basic programming skills"
  - "CS students"
required_courses: []
prerequisites:
  - "Basic python programming skills"
preparation: []
outcomes:
  - "Overview of a whole data science pipeline"
duration: 2 hours
difficulty: 4 # from 10
tags:
  - introduction
  - data cleaning
---
type: "LINK"
preview: "/media/courses/the_data_science_pipeline.png"
url: "https://prismia.chat/shared/data-science-pipeline"
```

TRY IT YOURSELF!

https://data-ai-literacy.ml

THANK YOU FOR YOUR ATTENTION!