Data Integration and Large Scale Analysis

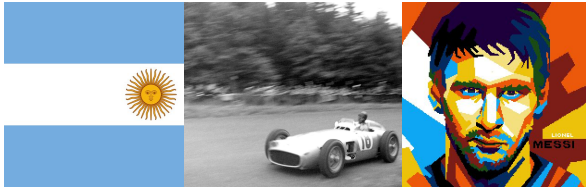
01- Introduction and Overview

Dr. Lucas Iacono - 2025
Know Center Research GmbH & Graz University of Technology

Agenda

- Team
- Organization
- Motivation and Goals
- Lecture Plan and Project

About Me



HCC
Institute of
HUMAN-CENTRED COMPUTING

How to contact me?

- liacono@tugraz.at
- lucasiacono
- SAL Building - 2nd Floor - Office 02 067
- https://lucasiacono.github.io



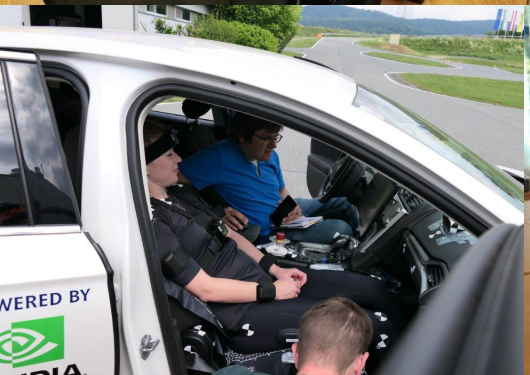

HCC
Institute of HUMAN-CENTRED COMPUTING

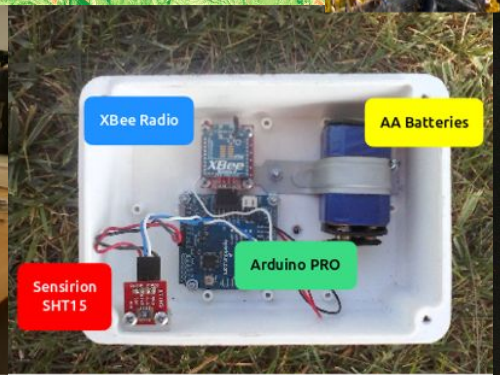
About Me

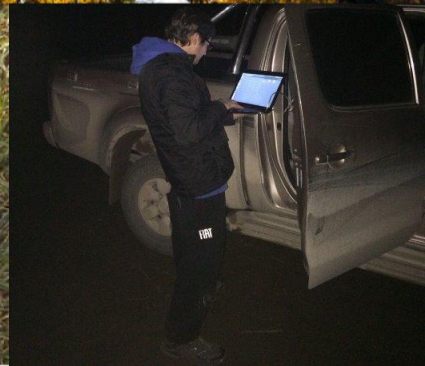
- **Degree in Electrical and Electronics Engineering** – University of Mendoza, Argentina (2007)
- **Doctor in Engineering (PhD)** - University of Mendoza, Argentina (2015)
- **Data Engineer** – Fiat Petronas PSG16 Race Team (2015 - 2016)
- **Associate Professor** -
  - Introduction to Technology – National University of Cuyo (2015 – 2019) - Argentina
  - Industrial Robotics – Computer Engineering – University of Mendoza (2008 – 2019) - Argentina
  - Postdoc – IoT Devices and UAVs applied to frost damage mitigation - Argentine Research Council (CONICET) - 2017
- **Know Center Research GmbH**
  - Senior Researcher - HAI Area (2019 - 2023)
  - Research Area Manager - Data Management for AI Area (2023 - Now)


playseat


MOPAR
35
PETRONAS
Seguros
Federal




XBee Radio
AA Batteries
Arduino PRO
Sensirion
SHT15



Data Management for AI @Know Center

Area Manager

Lucas Iacono

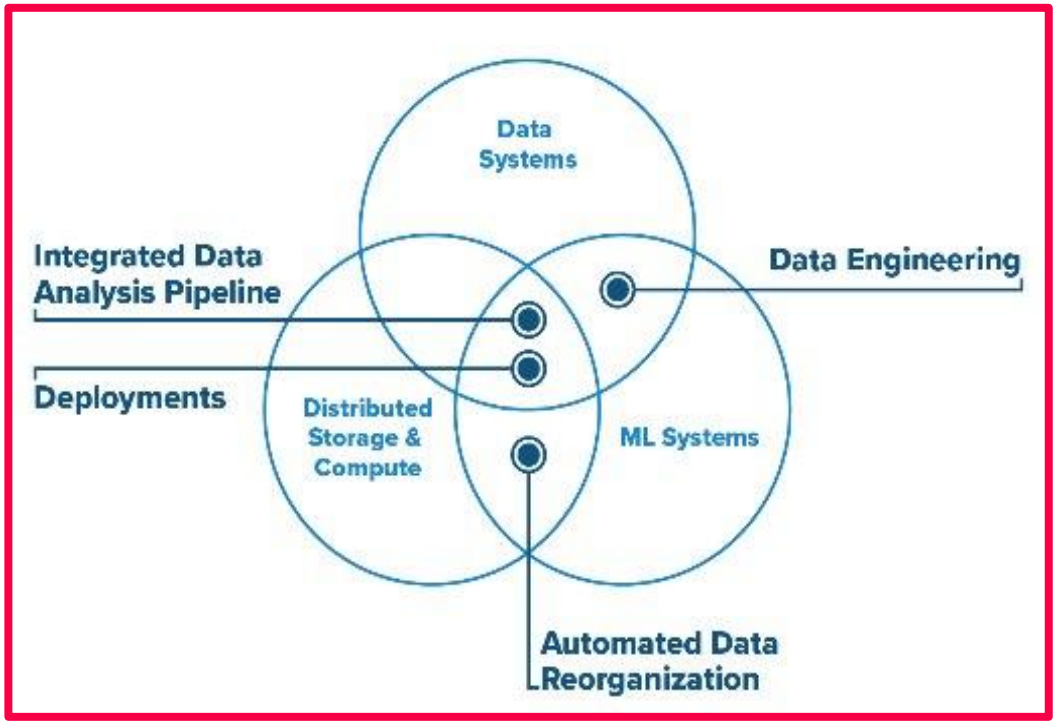
Researchers

Shafaq Siddiqi - Mark Dokter

Data Scientists

Alexander Hiebl - Lorenz Dirry - Snehila Snehila

Our research



HCC
Institute of
HUMAN-CENTRED COMPUTING

Course Organization

General Information

- **Title:** Data Integration and Large-Scale Analysis (DILA)
- **Type: VU** (means continuous assessment)
- **Semester hours:** 2 (3 ECTS)
- **Mandatory project:** 1 (2 ECTS)
- **Weekly lectures:** Fri 3:00 pm HS i5, attendance **optional**
- **Recommended papers** for additional reading on your own
- **Offered in:** WS

General Information

- **Studies:**
  - Bachelor program computer science (CS)
  - Bachelor program software engineering & management (SEM)
  - Master programs CS and SEM
    - Catalog Data Science: compulsory course in major/minor
  - Free subject course in any other study program or UNI

Prerequisites

- **Preferred:** Databases / Data Management
- **Sufficient:** basic understanding of SQL / Data Management / Relational Algebra
- **Basic programming skills** (**Python**, R)

General Information

- **Teach Center & Discord:**
  - Discord for questions and discussions: https://discord.gg/Xbk8MRVW
  - Teach Center for Learning Material
  - All classes will be recorded and uploaded to Tube
  - Google Colab for Class Exercises
  - E-mails only for important organizational questions!
- **Language and Communication:** English
  - Informal language (first name is fine)
  - Please also use English when asking questions
  - Ask for feedback (unclear content, missing background)
  - Online availability: Mo-Fri 8:00 to 17:00
  - Hello and Thanks when asking questions are welcome!
  - Office hours: by appointment or after lecture

Assessment

- **Mandatory Project (**100 points, **score to pass: 51 points):**
  - **Groups of 3**
  - **Develop a entity matching pipeline for the DBLP-Scholar Dataset**
  - **Task 01:** Entity Matching Pipeline (40 points)
  - **Task 02:** Feature Vector and ML Model (50 points)
  - **Task 03:** Reporting & Reproducibility (10 points)

Assessment

- **Mandatory Final Written Exam (**100 points, **score to pass: 51 points):**
  - **51 points:** needed for passing the exam
  - **Two dates:**30/01/2026 and 20/03/2026

- **Course Final Grade:**
  - Weighted grade between the project and the final exam. (0,3 * P + 0,7 * E)
  - Project: 30% of the final grade
  - Exam: 70% of the final grade.

- **Both the project and the final exam must be passed in order to pass the course.**

Assessment

- **Grading Scheme** (> 51 points needed **for passing the course**)
  - 0 - 50 points: 5
  - 51 - 60 points: 4
  - 61 - 74 points: 3
  - 75 - 90 points: 2
  - 91 - 100 points: 1

Course team

- **Lecturer:** Lucas Iacono

- **Student assistant (tutor):** Saiful Islam
  - s.islam@student.tugraz.at

Course Motivation

Data Sources and Heterogeneity

Some important concepts and keywords

Integration

Process of combining and harmonizing data from multiple sources into a unified, coherent format that can be put to use for various analytical, operational and decision-making purposes.

IBM

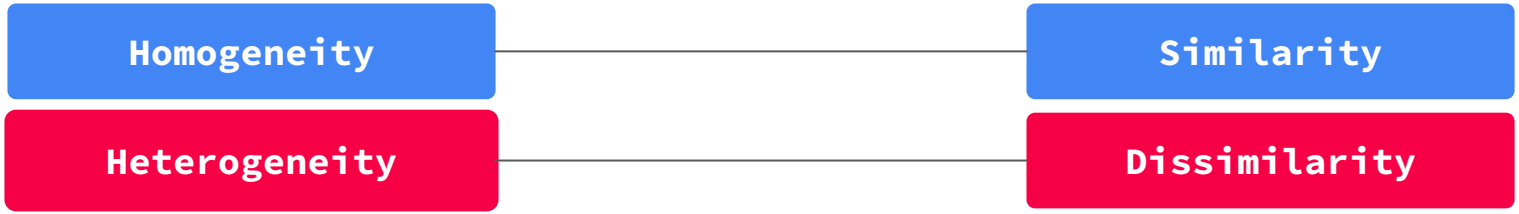
Problem of combining data residing at different sources, and providing the user with a unified view of these data.

Lenzerini, M. (2002, June). Data integration: A theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246).

HCC Institute of HUMAN-CENTRED COMPUTING

Data Sources and Heterogeneity

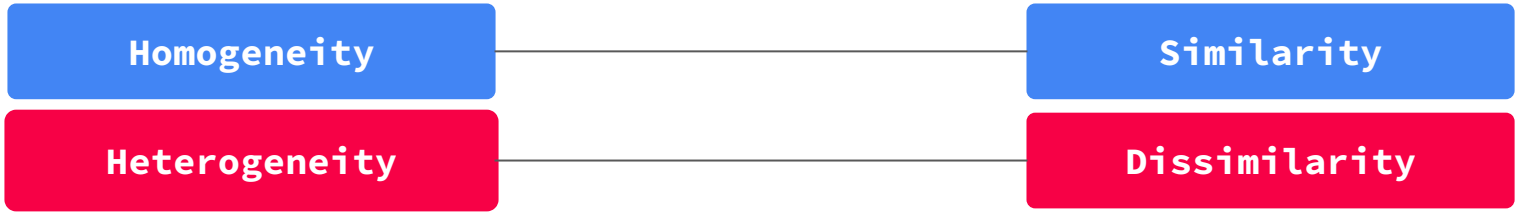
Some important concepts and keywords



Register	Name	Class	Engine	Format
A	Corvette C7	GT3	Petrol	CSV
B	Porsche 911	GT3	Petrol	HDF5

Data Sources and Heterogeneity

Some important concepts and keywords



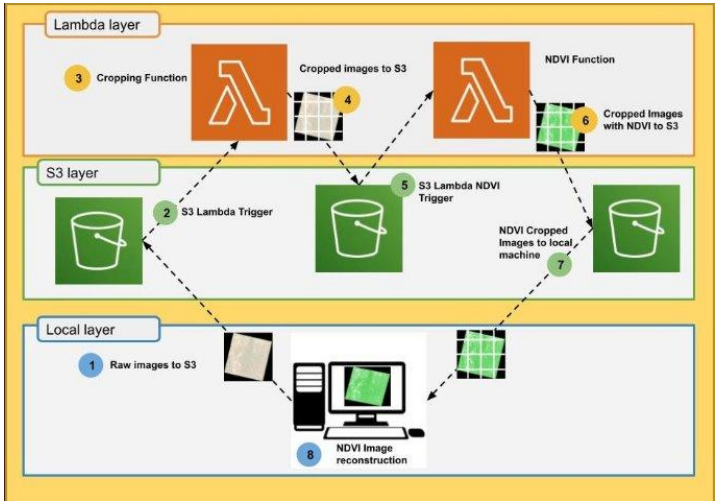
Register	Name	Class	Engine	Format
A	Corvette C7	GT3	Petrol	CSV
B	Porsche 911	GT3	Petrol	HDF5

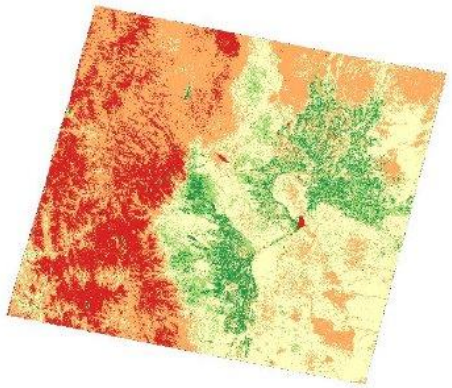
- **Semantic** ---> Meaning (GT3 petrol race cars)
- **Ontology** ---> Conceptual Structure (Name/Class/Engine)
- **Format** ---> Physical Representation (Binary vs Text)

HCC Institute of HUMAN-CENTRED COMPUTING

Data Sources and Heterogeneity: IT Infrastructure

Computing NDVI at the Cloud

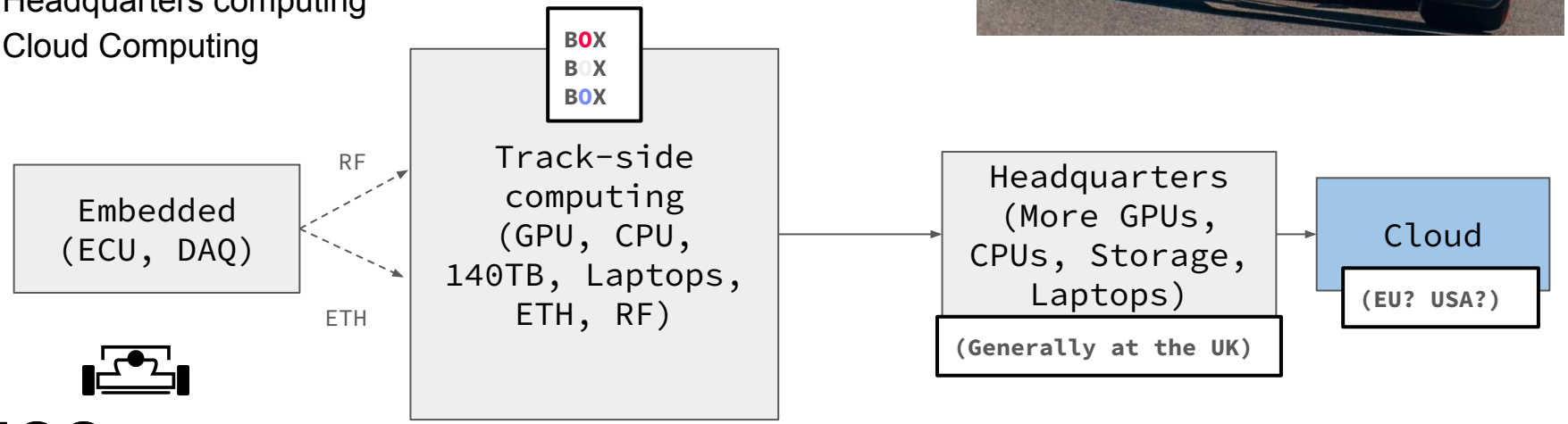




Iacono, L., Pacios, D., & Vázquez-Poletti, J. L. (2023). SNDVI: A new scalable serverless framework to compute NDVI. Frontiers in High Performance Computing, 1, 1151530.

Data Sources and Heterogeneity: Heterogeneous IT Infrastructure

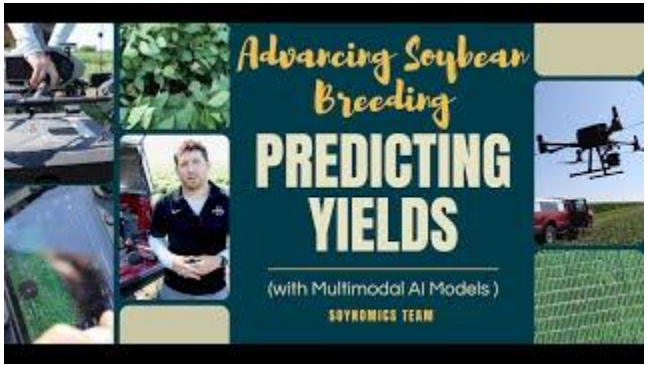
- ECU and Car Systems
- Telemetry
- Laptops
- Track-side computing
- Headquarters computing
- Cloud Computing



Multi-modal data (Agriculture)

Let's see if we can identify the heterogeneity

- Structure
- Format
- Type
- Storage

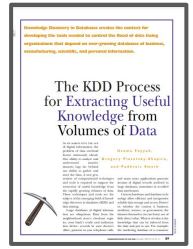


How we can integrate them?

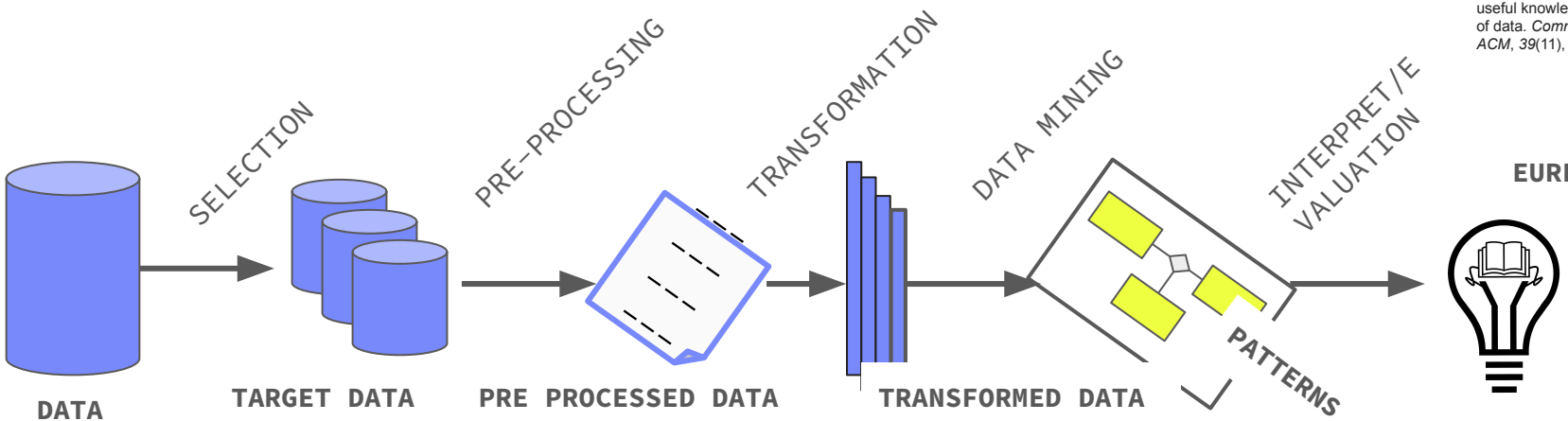
From data to models

From KDD to AI pipelines

- **Classic KDD** (Knowledge Discovery in Databases)
- Descriptive (association rules, clustering) and predictive

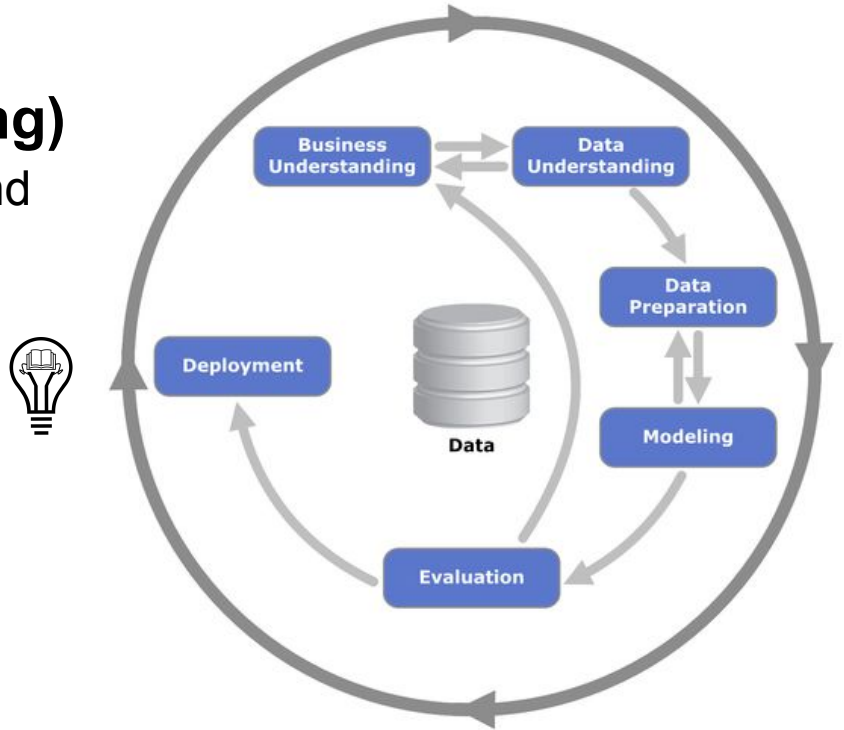


Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.



From KDD to AI pipelines

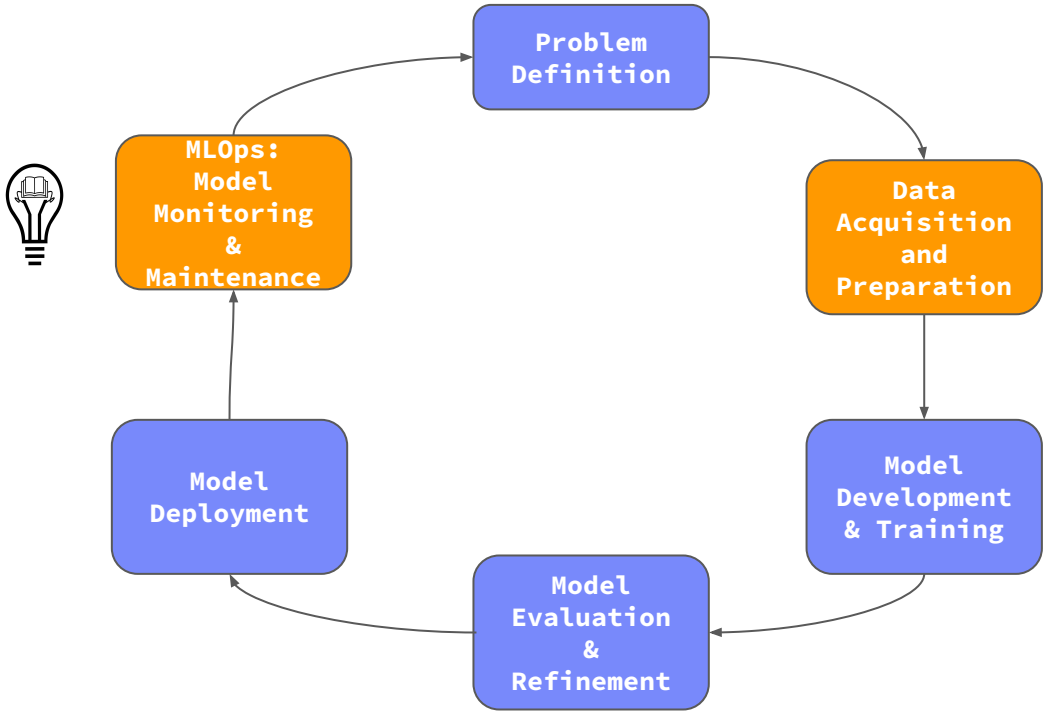
- **CRISP-DM (Cross-Industry Standard Process for Data Mining)**
- **What's new?** Business Understanding and Deployment **(A business perspective)**



Source: Statistik Dresden

From KDD to AI pipelines

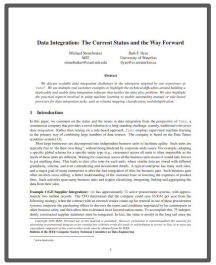
AI Lifecycle



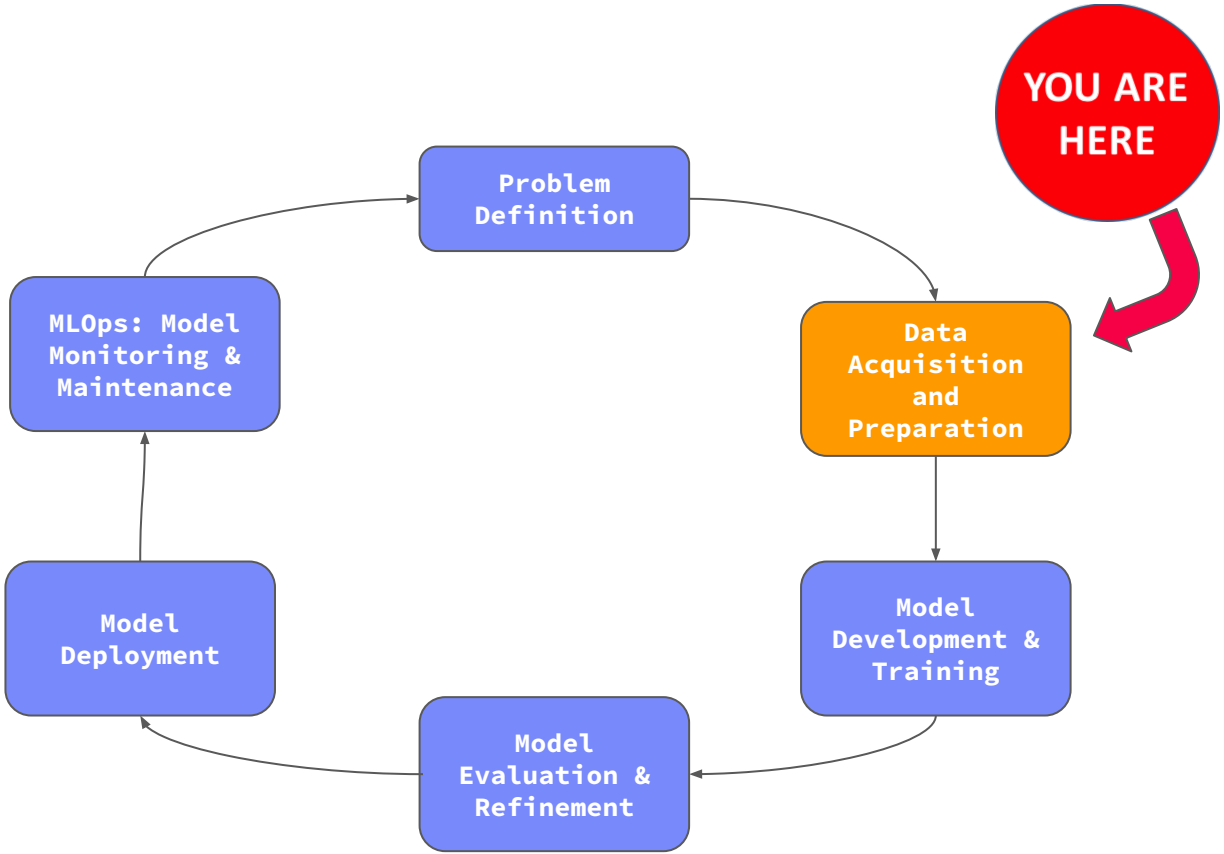
Ref: Jeffrey Saltz. Data Science Process Alliance.

The 80% Argument

Data scientists spend **80-90%** time on finding, integrating, cleaning datasets



Stonebraker, M., & Ilyas, I. F. (2018). Data Integration: The Current Status and the Way Forward. IEEE Data Eng. Bull., 41(2), 3-9.



Some open topics in AI pipelines

- **Multi-party data sharing.**
  - How to share data between organizations in multi-party and cross-country projects?
- **Data spaces technologies.**
  - Data spaces have been conceptualized to guarantee sovereign data sharing in multi-party environments by the signature of digital agreements for the use and distribution of data.

Some open topics in AI pipelines

- **Adaptability** Some of the most time-consuming processes in the AI lifecycle are detecting failures and decreases in data quality, such processes requires human intervention.
  - **Automated mechanisms** to predict and detect failures, avoiding the intervention of pipeline managers and flow guardians.
  - **Real-time data quality monitoring**. If there is a detection of a data inconsistency, the DQM module can request data resubmission, avoiding the involvement of humans.

Some open topics in AI pipelines

- **Improvement of data traceability.** Some components of the AI lifecycle can experience errors. These errors can lead to data loss, and it is difficult to identify at which point in the pipeline the error occurred.
    - **Pipelines for AI need to automatically track the quality and reliability of the data and metadata obtained in each of their processes.**

Course Goals

By the end of this course, you will be able to:

- Understand major data integration architectures and their role in modern data ecosystems.
- Apply key techniques for data integration and cleaning to ensure consistency, quality, and usability of data.
- Evaluate methods for large-scale data storage and analysis, with a focus on scalability and efficiency.

Course Calendar

Part A: Data Integration and Preparation

- Data Integration Architectures
  - October 03. Introduction and Overview
  - October 10. Data Warehousing, ETL, and SQL/OLAP
  - October 17. Message-oriented Middleware, EAI, and Replication + **Project Presentation**
- Key Integration Techniques
  - October 24. Schema Matching and Mapping
  - October 31. Entity Linking and Deduplication
  - **November 7. No lecture (Time for project preparation)**
  - November 14. Data Cleaning and Data Fusion

Part B: Large-Scale Data Management & Analysis

- Cloud Computing
  - November 21. Cloud Computing Fundamentals
  - November 28. Cloud Resource Management and Scheduling
  - December 05. Distributed Data Storage
- Large-Scale Data Analysis
  - December 12. Distributed, Data-Parallel Computation
  - December 19. Distributed Stream Processing
  - **January 09. No Lecture (Time to prepare the project submission)**
  - January 16. Distributed Machine Learning Systems
  - **January 23. No Lecture (Time to prepare for the exam)**

Project and Exams

- Project
  - January 10. **Project Submission Deadline**
- Exam
  - January 30th (First Exam)
  - March 20th (Second Exam)

Summary and Q&A

Summary and Q&A

- **Course Goals**
  - Major data integration architectures
  - Key techniques for data integration and cleaning
  - Methods for large-scale data storage and analysis
- **Next Lectures**
  - Data Warehousing, ETL, and SQL/OLAP [Oct 10]
  - Message-oriented Middleware, EAI, and Replication [Oct 17]

Many thanks!