Data Integration and Large Scale Analysis

Slides credit: Prof. Matthias Boehm, Dr. Shafaq Siddiqi

01- Introduction and Overview

Lucas Iacono. PhD. - 2024

TU Graz

SCIENCE
PASSION
TECHNOLOGY

ISDS

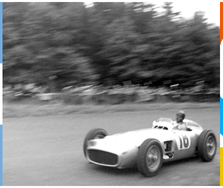
How to contact me?

- liacono@know-center.at (for now…)
- lucasiacono
- SAL Building - 2nd Floor - Office 02 067
- https://lucasiacono.github.io/ws2024_2025_dia.html

Agenda

- Organization
- Motivation and Goals
- Learning Units and Projects

About Me



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)
- **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)

Data Management for AI @Know Center

Area Manager

Lucas Iacono

Researchers

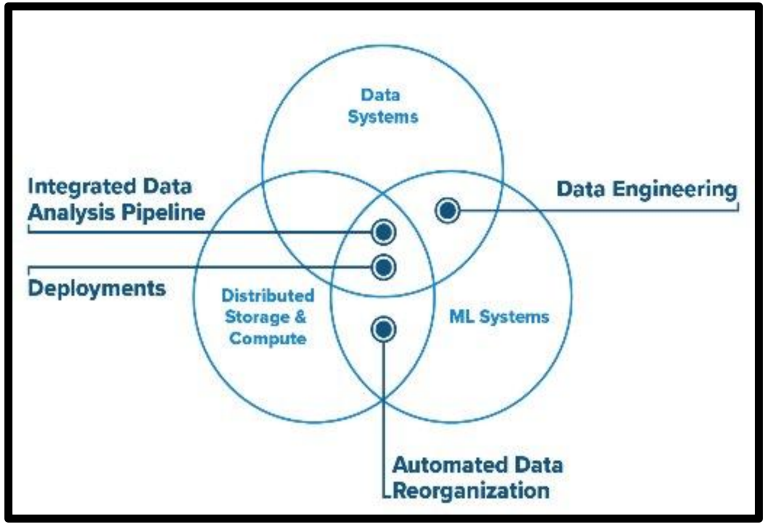
Shafaq Siddiqi - Mark Dokter

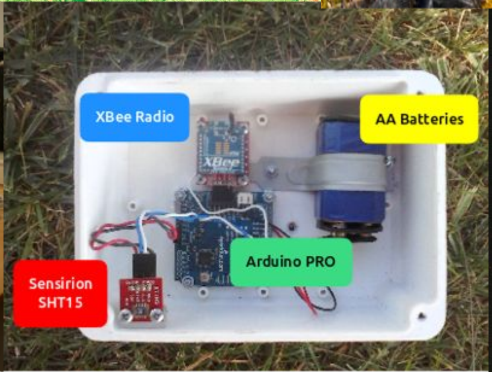
Data Scientists

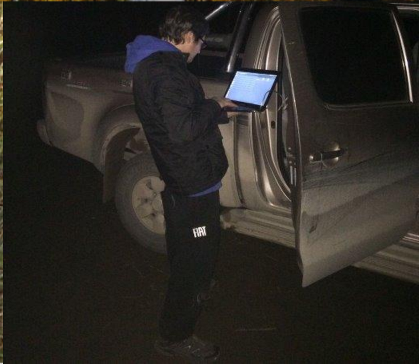
Alexander Hiebl - Lorenz Dirry

Our research

- **Data Platforms and Cloud**
  - Adapt and tune data management and ML systems for domain-specific data platforms, federated learning, and hybrid cloud-edge.
- **Integrated Data Analysis Pipelines**
  - Simplify the full utilization of heterogeneous hardware, and adopt this infrastructure in related multi-firm projects
- **Automatic Data Reorganization**
  - Address the increasing redundancy in complex data science workflows for data preparation and cleaning, data augmentation, feature engineering, hyper-parameter optimization, and model training.


Data Systems
Integrated Data Analysis Pipeline
Data Engineering
Deployments
Distributed Storage & Compute
ML Systems
Automated Data Reorganization


XBee Radio
AA Batteries
Arduino PRO
Sensirion SHT15



Course Organization

Basic Course Organization

- **Team**
  - Lecturers:
    - Dr. M.Sc. Shafaq Siddiqi, ISDS - Know Center Research
    - Lucas Iacono PhD., ISDS - Know Center Research
  - Tutor:
    - Saiful Islam, ISDS
- **Language**
  - Classes and slides: **English**
  - Communication and examination: **English**

Basic Course Organization

- **Course Format**
  - **VU 2/1**, **5 ECTS** (2 x 1.5 ECTS + 1x2 ECTS), bachelor/master (+info next class)
  - **Weekly lectures** (Fri 3pm, HS i5 + Webex, including Q&A), attendance **optional**
  - **Mandatory exercises or programming project** (2 ECTS)
  - **Recommended papers** for additional reading on your own
- **Prerequisites**
  - **Preferred:** course Data Management / Databases (a very good boot system)
  - **Sufficient:** basic understanding of SQL / Data Management / Relational Algebra
  - **Basic programming skills** (**Python**, Java, R)

Basic Course Organization

- Video Recording
    - Link in **TUbe & TeachCenter**
    - **Optional** attendance
    - **Hybrid**, in-person but video-recorded lectures
    - **HS i5** and **Webex**:
        - https://tugraz.webex.com/meet/shafaq.siddiqi (1st module)
        - https://tugraz.webex.com/meet/lucas.iacono (2nd module)

and more...course organization

- Website
  - https://lucasiacono.github.io/ws2024_2025_dia.html
  - All course material (lecture slides) and dates
- Video Recording Lectures (TUbe)
- Communication
  - Informal language (Lucas and Shafaq)
  - Ask for feedback!!! (unclear content, missing background)
  - Newsgroup: N/A - email is fine, TeachCenter forum for discussions
  - Office hours: by appointment or after lecture
- Exam
  - **Completed exercises or project**
  - **Final written exam** (oral exam if <25 students take the exam and for Erasmus students)
  - **Grading** (30% project/exercises completion, 70% exam)

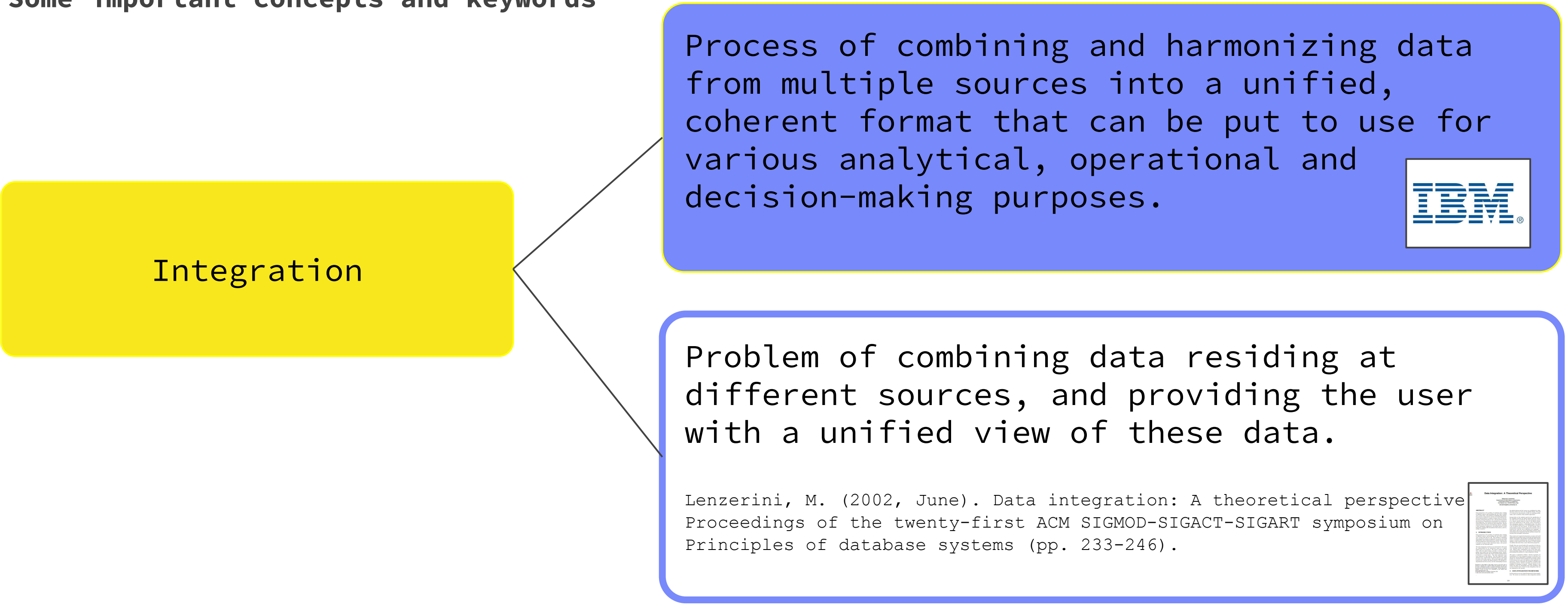
and more...more...course organization

- **Course Applicability**
  - **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

Course Motivation and Goals

Data Sources and Heterogeneity

Some important concepts and keywords



Integration

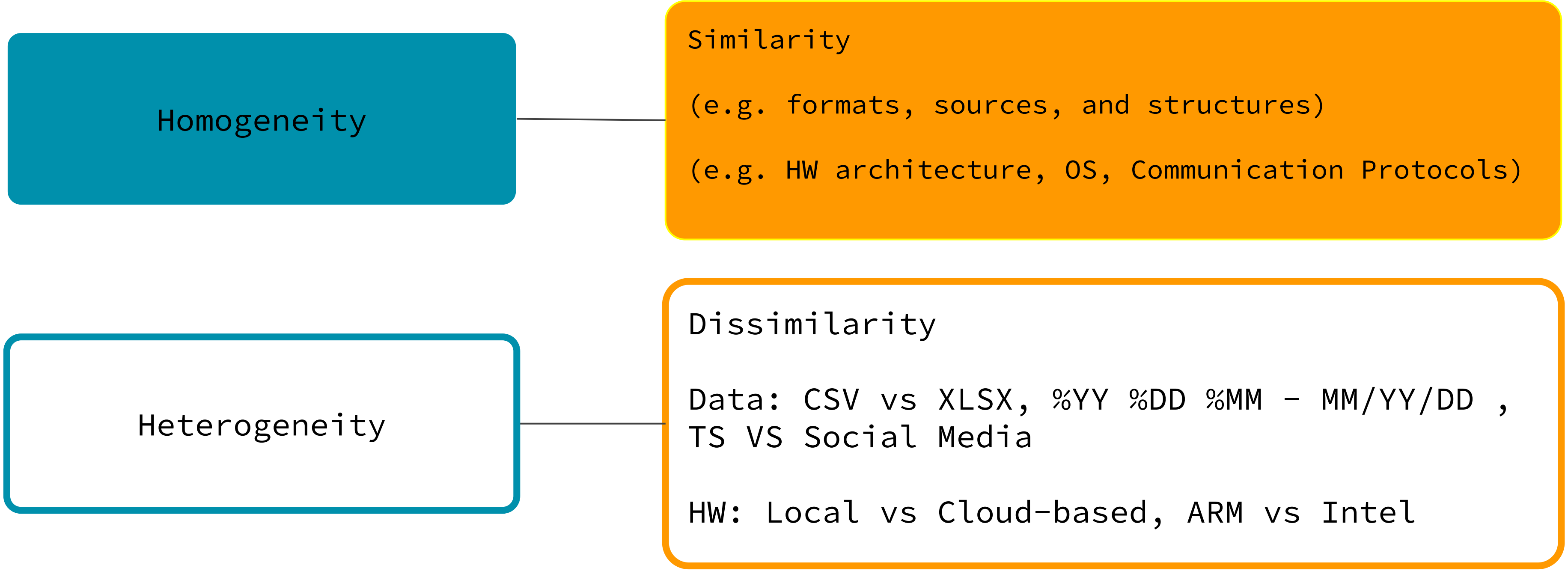
Process of combining and harmonizing data from multiple sources into a unified, coherent format that can be put to use for various analytical, operational and decision-making purposes.

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).

Data Sources and Heterogeneity

Some important concepts and keywords



Homogeneity

Similarity

(e.g. formats, sources, and structures)

(e.g. HW architecture, OS, Communication Protocols)

Heterogeneity

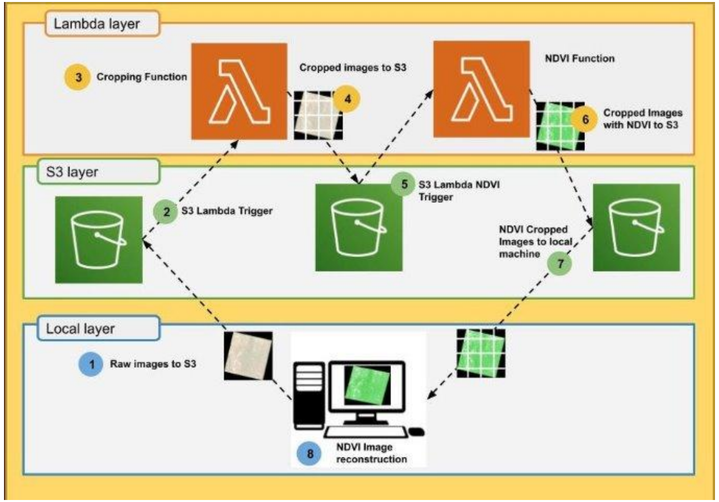
Dissimilarity

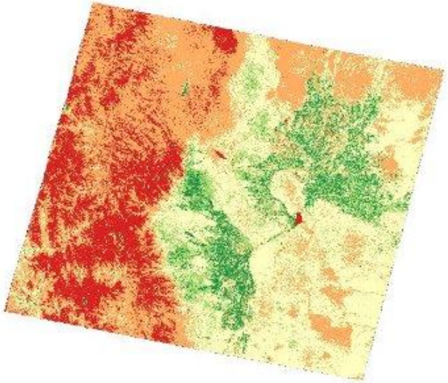
Data: CSV vs XLSX, %YY %DD %MM - MM/YY/DD , TS VS Social Media

HW: Local vs Cloud-based, ARM vs Intel

Data Sources and Heterogeneity: IT Infrastructure

Computing NDVI at the Cloud

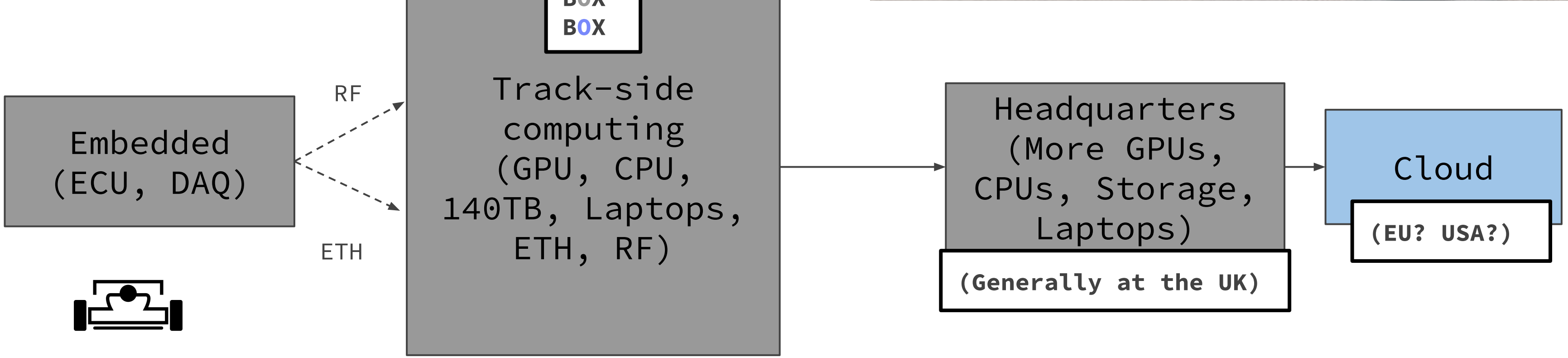




Iacono, L., Pacios, D., & Vázquez-Poletti, J. L. (2023). SNDVI: A new scalable serverless framework to compute NDVI. Frontiers in High Performance Computing, 1, 1151530.

Data Sources and Heterogeneity: Heterogeneous IT Infrastructure

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


BOX
BOX
BOX
Embedded (ECU, DAQ)
RF
ETH
Track-side computing (GPU, CPU, 140TB, Laptops, ETH, RF)
Headquarters (More GPUs, CPUs, Storage, Laptops)
(Generally at the UK)
Cloud
(EU? USA?)

Multi-modal data (Agriculture)

Let's see if we can identify the heterogeneity

- Structure
- Format
- Type
- Storage

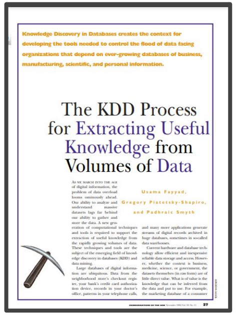
How we can integrate them?

From data to models

From KDD to the AI Lifecycle

Classic KDD (Knowledge Discovery in Databases)

Descriptive (association rules, clustering) and **predictive**



Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

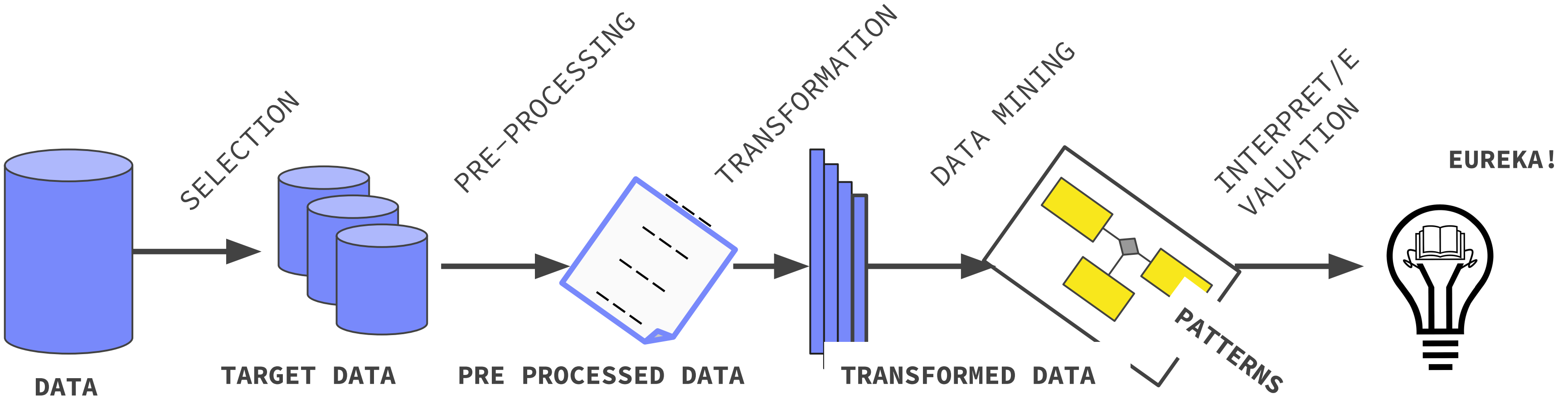
As we march into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

COMMUNICATIONS OF THE ACM November 1996/Vol. 39, No. 11 27

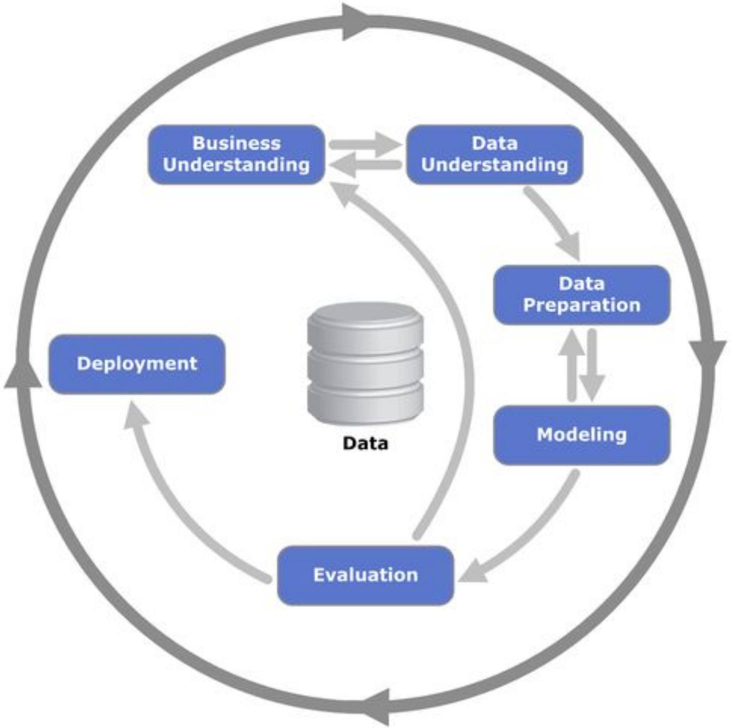
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.



From KDD to the AI Lifecycle

CRISP-DM (Cross-Industry Standard Process for Data Mining)

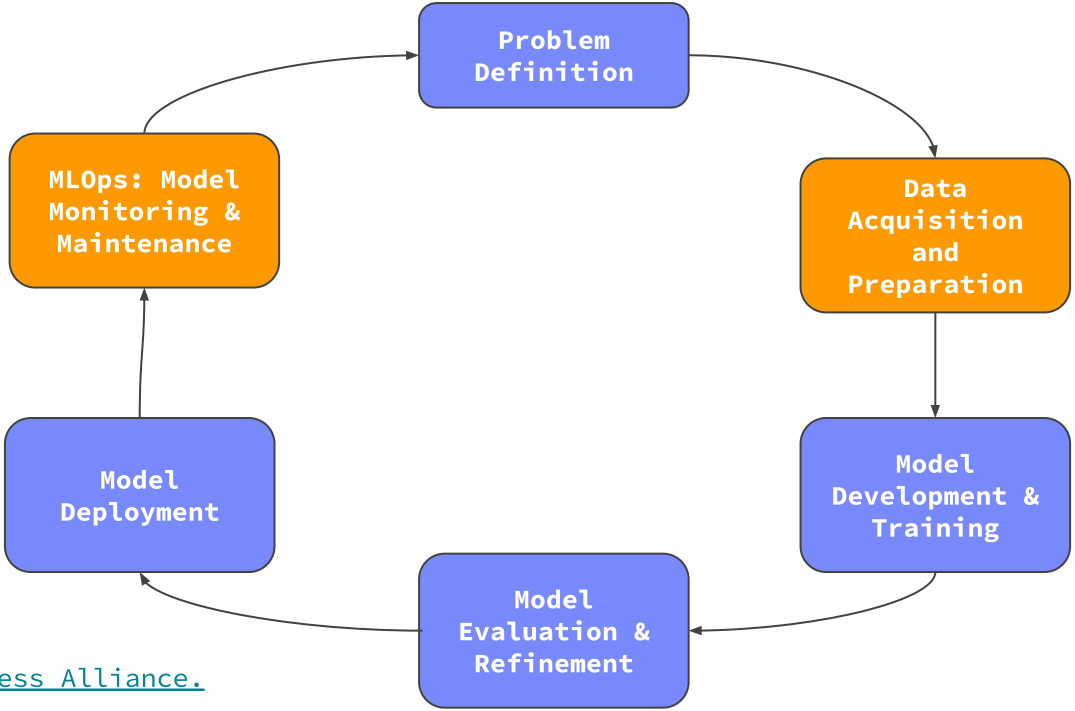
What's new? Business Understanding and Deployment **(A business perspective)**



Source: Statistik Dresden

From KDD to the AI Lifecycle

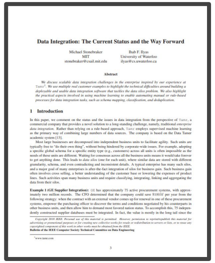
AI Lifecycle


Problem Definition
Data Acquisition and Preparation
Model Development & Training
Model Evaluation & Refinement
Model Deployment
MLOps: Model Monitoring & Maintenance

Ref: Jeffrey Saltz. Data Science Process Alliance.

The 80% Argument

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



Data Integration: The Current Status and the Way Forward

Michael Stonebraker
MIT
stonebraker@csail.mit.edu

Ihab F. Ilyas
University of Waterloo
ilyas@cs.uwaterloo.ca

Abstract

We discuss scalable data integration challenges in the enterprise inspired by our experience at Tamr[1]. *We use multiple real customer examples to highlight the technical difficulties around building a deployable and usable data integration software that tackles the data silos problem. We also highlight the practical aspects involved in using machine learning to enable automating manual or rule-based processes for data integration tasks, such as schema mapping, classification, and deduplication.*

1 Introduction

In this paper, we comment on the status and the issues in data integration from the perspective of Tamr, a commercial company that provides a novel solution to a long-standing challenge, namely, traditional *enterprise data integration*. Rather than relying on a rule-based approach, Tamr employs supervised machine learning as the primary way of combining large numbers of data sources. The company is based on the Data Tamer academic system [13].

Most large businesses are decomposed into independent business units to facilitate agility. Such units are typically free to "do their own thing", without being hindered by corporate-wide issues. For example, adopting a specific global schema for a specific entity type (e.g., customers) across all units is often impossible as the needs of these units are different. Waiting for consensus across all the business units means it would take forever to get anything done. This leads to *data silos* (one for each unit), where similar data are stored with different granularity, schema, and even contradicting and inconsistent details. A typical enterprise has many such silos, and a major goal of many enterprises is after-the-fact integration of silos for business gain. Such business gain often involves cross selling, a better understanding of the customer base or lowering the expenses of product lines. Such activities span many business units and require classifying, integrating, linking and aggregating the data from their silos.

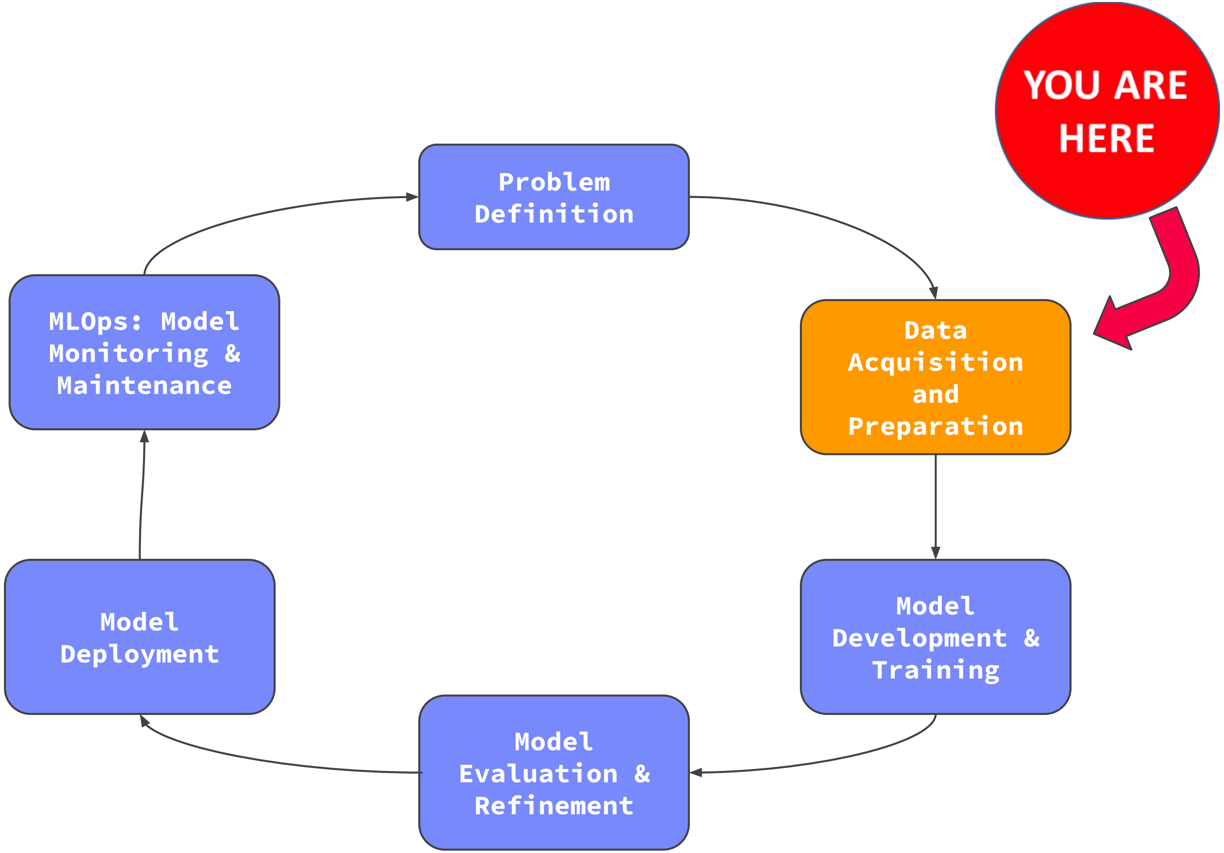
Example 1 (GE Supplier Integration): GE has approximately 75 active procurement systems, with approximately two million records. The CFO determined that the company could save $100M per year from the following strategy: when the contract with an external vendor comes up for renewal in one of these procurement systems, empower the purchasing officer to discover the terms and conditions negotiated by his counterparts in other business units, and then allow him to demand most favored nation status. To accomplish this, 75 independently constructed supplier databases must be integrated. In fact, the value is mostly in the long tail since the

Copyright 2018 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[1] www.tamr.com

3

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


YOU ARE HERE
Problem Definition
Data Acquisition and Preparation
Model Development & Training
Model Evaluation & Refinement
Model Deployment
MLOps: Model Monitoring & Maintenance

Some open topics in the AI lifecycle

- **Multi-party data sharing.**
    - How to share data between organizations in multi-party and cross-country projects?
- **Dataspaces 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 the AI lifecycle

- **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 the AI lifecycle

- **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

- Major data integration architectures
- Key techniques for data integration and cleaning
- Methods for large-scale data storage and analysis

Learning Units and Projects

Part A

Data Integration and Preparation

- LU1. Data Integration Architectures
  - Introduction and Overview [Oct 11]
  - Data Warehousing, ETL, and SQL/OLAP [Oct 18]
  - Message-oriented Middleware, EAI, and Replication [Oct 25]

Part A

Data Integration and Preparation

- LU2. Key Integration Techniques
  - Schema Matching and Mapping [Nov 08]
  - Entity Linking and Deduplication [Nov 15]
  - Data Cleaning and Data Fusion [Nov 22]

Part B

Large-Scale Data Management & Analysis

- LU3. Cloud Computing
  - Cloud Computing Fundamentals [Nov 29]
  - Cloud Resource Management and Scheduling [Dec 06]
  - Distributed Data Storage[Dec 13]

Part B

Large-Scale Data Management & Analysis

- LU4. Large-Scale Data Analysis
  - Distributed, Data-Parallel Computation [Jan 10]
  - Distributed Stream Processing [Jan 17]
  - Distributed Machine Learning Systems [Jan 24]

Overview Projects or Exercises

Overview Projects or Exercises

- **Team**
  - **1 – 3 person teams (w/clearly separated responsibilities)**
- **Objectives**
  - Non-trivial programming project in DIA context **(2 ECTS 50 hours)**
  - **Exercise:** Data engineering and ML pipeline
    - Data cleaning and integration of multi-modal data sources
    - ML model training and evaluation
  - **Optional:** Open source contribution to Apache SystemDS or DAPHNE project **(from HW to high-level scripting)**
    - https://github.com/apache/systemds
    - https://github.com/daphne-eu/daphne
- **Timeline**
  - Oct 25: Exercise Description
  - Jan 17: Final project/exercise deadline

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 18]
  - Message-oriented Middleware, EAI, and Replication [Oct 25]

Vielen Dank!