Slides credit: Matthias Boehm

SCIENCE
PASSION
TECHNOLOGY

Data Integration and Large Scale Analysis
01 Introduction and Overview

Shafaq Siddiqi

Graz University of Technology, Austria

Last update: Oct 06, 2023

ISDS

Announcements/Org

- **#1 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

Agenda

- **Course Organization**
- **Course Motivation and Goals**
- **Course Outline and Projects**

About Me

- **09/2019 TU Graz**, Austria
  - Teaching Assistant, TU Graz
  - **Institute of Interactive Systems and Data Science, CSBME**
    (ML systems internals, end-to-end data science lifecycle)
    https://github.com/apache/systemds

- **2017-2019 Sukkur IBA University**
  - Lecturer (Computer Science)
  - Teaching and supervising FYPs in Bachelor programs

- **2020 PhD Student TU Graz**, Austria
  - Data preprocessing for Heterogeneous Large Scale Data
  - Generation and Optimization of Data Cleaning Pipelines

Data Management group

Course Organization

Basic Course Organization

- **Staff**
  - Lecturer: M.Sc. Shafaq Siddiqi, ISDS
- **Language**
  - Lectures and slides: **English**
  - Communication and examination: **English**
- **Course Format**
  - VU 2/1, **5 ECTS** (2x 1.5 ECTS + 1x 2 ECTS), bachelor/master
  - **Weekly lectures** (**Fri 3pm**, 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 is very good start
  - **Sufficient:** basic understanding of SQL / RA (or willingness to fill gaps)
  - Basic programming skills (Python, R, Java, C++)

Course Logistics

- **Website**
  - https://shafaq-siddiqi.github.io./dia2023.html
  - All course material (lecture slides) and dates
- **Video Recording Lectures (TUbe)**
- **Communication**
  - **Informal language** (first name is fine)
  - Please, **immediate 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)

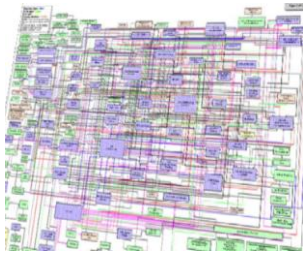
Course Logistics, cont.

- **Course Applicability**
  - **Bachelor** programs computer science (CS), as well as software engineering and management (SEM)
  - **Master** programs computer science (CS), as well as software engineering and management (SEM)
    - Catalog Data Science: **compulsory** course in major/minor
  - **Free subject course** in any other study program or university

Course Motivation and Goals

Data Sources and Heterogeneity

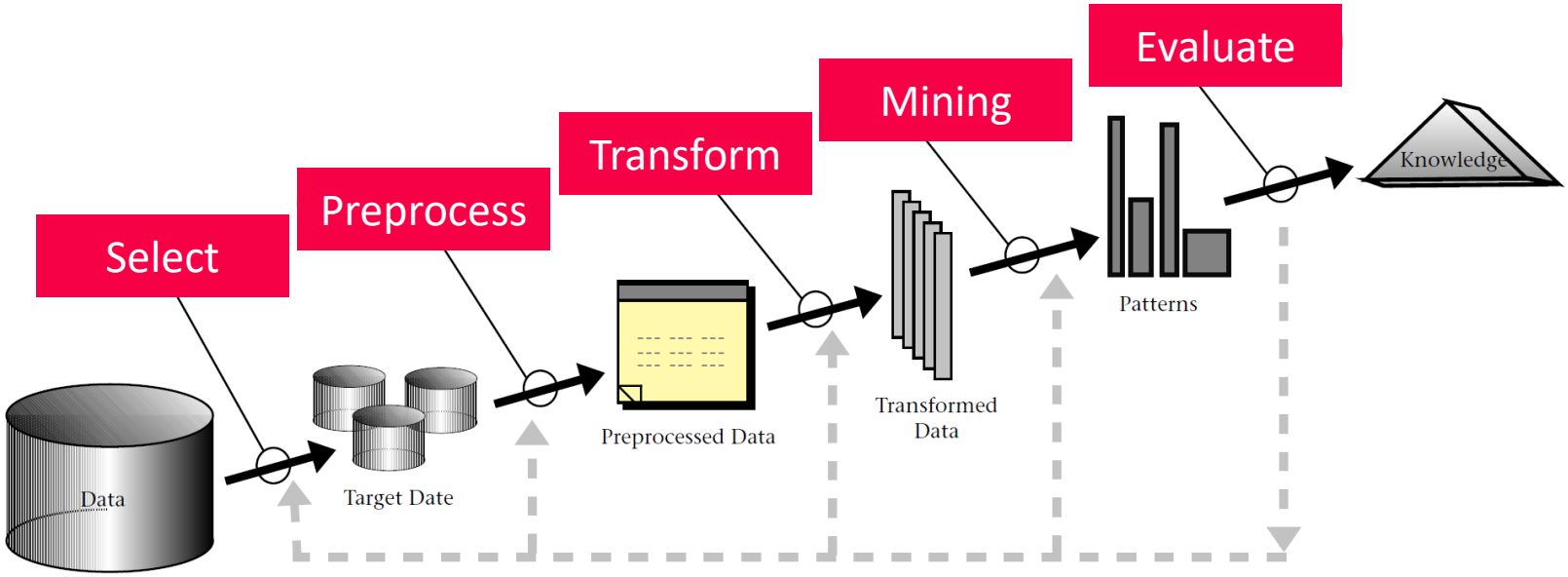
- **Terminology**
  - **Integration** (Latin integer = whole): consolidation of data objects / sources
  - **Homogeneity** (Greek homo/homoios = same): similarity
  - **Heterogeneity**: dissimilarity, different representation / meaning
- **Heterogeneous IT Infrastructure**
  - Common enterprise IT infrastructure contains >100s of **heterogeneous and distributed systems and applications**
  - E.g., health care data management: 20 - 120 systems



- **Multi-Modal Data (example health care)**
  - Structured patient data, patient records incl. prescribed drugs
  - Knowledge base drug APIs (active pharmaceutical ingredients) + interactions
  - Doctor notes (text), diagnostic codes, outcomes
  - Radiology images (e.g., MRI scans), patient videos
  - Time series (e.g., EEG, ECoG, heart rate, blood pressure)

The Data Science Lifecycle

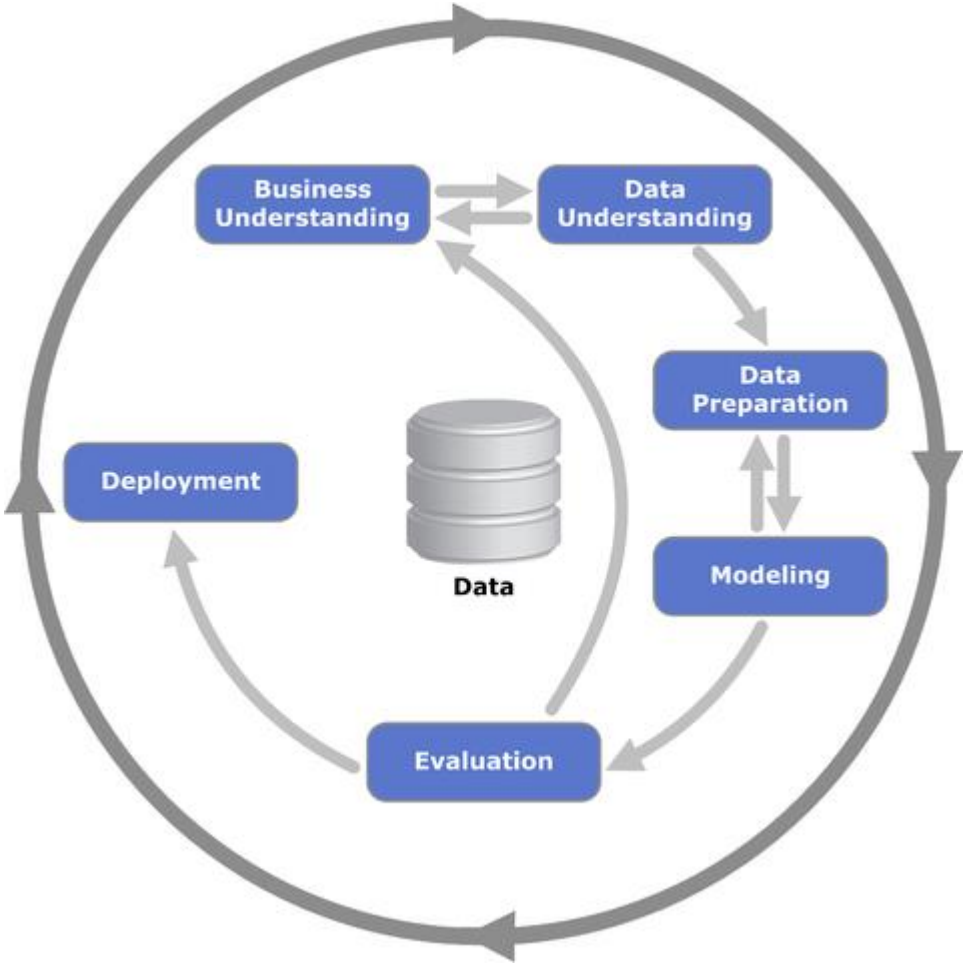
- **Classic KDD Process** (Knowledge Discovery in Databases)
  - Descriptive (association rules, clustering) and predictive



[Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth: From Data Mining to Knowledge Discovery in Databases. **AI Magazine 17(3) (1996)**]

The Data Science Lifecycle, cont.

- **CRISP-DM**
  - **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
  - Additional focus on business understanding and deployment



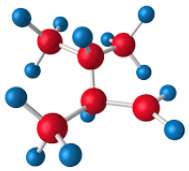
[https://statistik-dresden.de/archives/1128]

The Data Science Lifecycle, cont.

Data-centric View:
Application perspective
Workload perspective
System perspective

Data extraction, schema alignment, entity resolution, data validation, data cleaning, outlier detection, missing value imputation, semantic type detection, data augmentation, feature selection, feature engineering, feature transformations

Data Scientist



Data Integration
Data Cleaning
Data Preparation

Model Selection
Training
Hyper-parameters

Validate & Debug
Deployment
Scoring & Feedback

Data/SW Engineer

Exploratory Process
(experimentation, refinements, ML pipelines)

DevOps Engineer

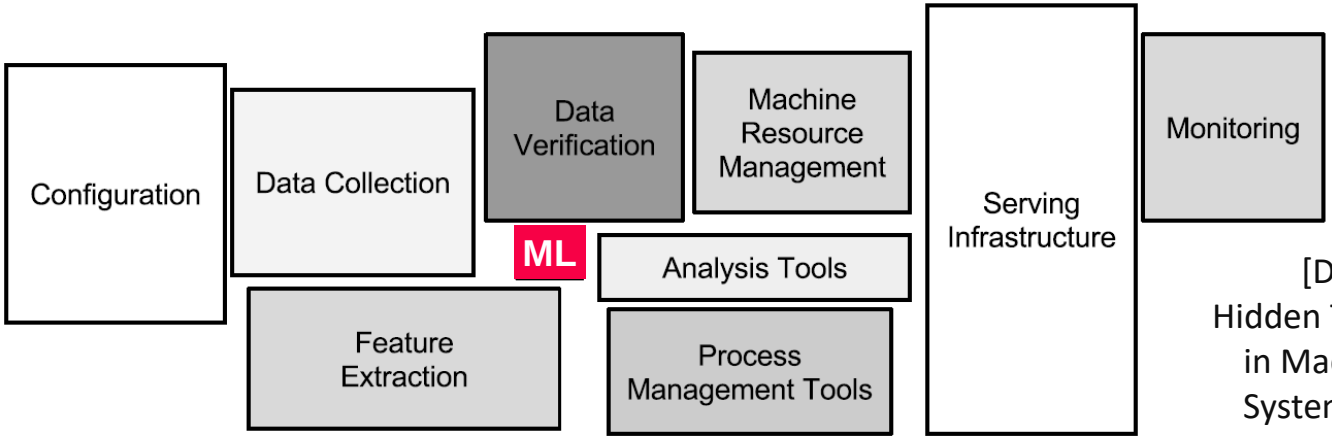
Key observation: SotA data integration/cleaning based on ML

The 80% Argument

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

[Michael Stonebraker, Ihab F. Ilyas: Data Integration: The Current Status and the Way Forward. **IEEE Data Eng. Bull. 41(2) (2018)**]

- **Technical Debts in ML Systems**

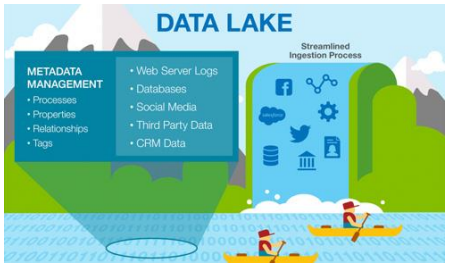


[D. Sculley et al.: Hidden Technical Debt in Machine Learning Systems. **NIPS 2015**]

  - Glue code, pipeline jungles, dead code paths
  - Plain-old-data types (arrays), multiple languages, prototypes
  - Abstraction and configuration debts
  - Data testing, reproducibility, process management, and cultural debts

Complementary Architectures

Distributed Computation Frameworks



Audio, Image, Video, Text, Streams, Logs

#1 Information System Pyramid

#2 Data Lake

Vertical Integration (e.g., ETL)

Catalogs

Distributed Data Stores

DSS

Strategic Systems

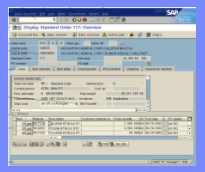
DWH

Analytical Systems

SCM

Material

ERP



CRM

eCommerce

Operational Systems

Horizontal Integration (e.g., EAI)

Course Goals

- **#1** Major data integration architectures
- **#2** Key techniques for data integration and cleaning
- **#3** Methods for large-scale data storage and analysis

Course Outline and Projects

Part A: Data Integration and Preparation

Data Integration Architectures

- **01 Introduction and Overview** [Oct 06]
- **02 Data Warehousing, ETL, and SQL/OLAP** [Oct 13]
- **03 Message-oriented Middleware, EAI, and Replication** [Oct 20]

Key Integration Techniques

- **04 Schema Matching and Mapping** [Oct 27]
- **05 Entity Linking and Deduplication** [Nov 03]
- **06 Data Cleaning and Data Fusion** [Nov 10]

Part B: Large-Scale Data Management & Analysis

Cloud Computing

- **07 Cloud Computing Foundations** [Nov 17]
- **08 Cloud Resource Management and Scheduling** [Nov 24]
- **09 Distributed Data Storage** [Dec 01]

Large-Scale Data Analysis

- **10 Distributed, Data-Parallel Computation** [Dec 15]
- **11 Distributed Stream Processing** [Jan 12]
- **12 Distributed Machine Learning Systems** [Jan 19]

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** https://github.com/apache/systemds (from HW to high-level scripting)
- **Timeline**
  - **Oct 20:** Exercise description
  - **Jan 12:** Final project/exercise deadline

Summary and Q&A

- **Course Goals**
  - **#1** Major data integration architectures
  - **#2** Key techniques for data integration and cleaning
  - **#3** Methods for large-scale data storage and analysis

- **Next Lectures**
  - 02 **Data Warehousing, ETL, and SQL/OLAP** [Oct 13]
  - 03 **Message-oriented Middleware, EAI, and Replication** [Oct 20]