# Einführung in MLOps

## 21 Data Validation

Tobias Mérinat
teaching2025@fsck.ch

Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN

Department of Information Technology
Lucerne University of Applied Sciences and Arts
6343 Rotkreuz, Switzerland

14. und 15. Februar 2025

# Datenvalidierung

- Validierung des Schemas der Input-Daten
- Rein syntaktisch
  - nicht semantisch
  - nicht statistisch

# Motivation und Ziel

- Eine der häufigsten Ursachen von Fehlern in Datenpipelines
- Ziele
  - frühzeitiges Erkennen von Datenproblemen
  - möglichst aussagekräftige Fehlermeldungen

# Typische Tests

- Anzahl Spalten, Spalten-Namen
- Spalten-Reihenfolge
- Datentypen
- Daten-Ranges (min, max oder Werte-Set)
- Regexps
- Not Null
- Key Constraints

# Tools für die Datenvalidierung

- PyDantic
  - Verbreitet, sehr Pythonic
  - Fokus auf Klassen, nicht DataFrames
  - Nicht gut geeignet für Data Pipelines
- Pandera
  - Unterstützt Pandas, PySpark, Polars
  - Relativ flache Lernkurve
- Great Expectations
  - DeFacto Tool für produktive Datenvalidierung
  - On-Premise und Cloud Lösung
  - Komplettes System mit Reports, Profiling, Documentation Generation und Tool-Integration (Airflow, dbt, Spark, SQL Datenbanken, Slack, . . . )
  - Steile Lernkurve

Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN