DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# Formal Description of Validity Criteria for Data and Automated Cloud-Based Validity Checking to Generate a Validated Data Set

**Lucas Krauße**

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# Formal Description of Validity Criteria for Data and Automated Cloud-Based Validity Checking to Generate a Validated Data Set

# Formale Beschreibung von Gültigkeitskriterien für Daten und automatisierte cloud-basierte Gültigkeitsprüfung zur Erzeugung einer validierten Datenmenge

| | |
|---|---|
| Author: | Lucas Krauße |
| Supervisor: | Prof. Dr.-Ing. Bhatotia Pramod |
| Advisor: | Christian Vetter |
| Submission Date: | 12.08.2022 |

# Abstract

**Fa**hrzeugbetriebs und **S**ervicedaten **T**ransfer und **A**nalyse (FASTA) data are user-related vehicle data that are used to improve the quality of BMW vehicles and to further develop future engines. So far, the data are used for analyses and evaluations without uniform and formal validation. This carries the risk of errors and duplication of effort of validation approaches. To counteract this problem, this thesis is about validating the raw data with formal validation rules on known anomalies by using the validation framework Pandera [1]. To improve the creation, extension, adaptation, and communication of the formal validation rules, a **D**omain **S**pecific **L**anguage (DSL), called *Conditional Language*, is created. It is a propositional logic DSL built on formal methods for checking Panda's [2] data structures against a user-defined logic sequence, including arithmetic operations.
The general implementation is based on creating a pipeline that first loads the raw FASTA data, then validates it with Pandera combined with the *Conditional Language*, and flagging the data based on violations regarding the defined validation rules for future analyses and evaluations. To validate the large amount of histogram data, two classification approaches are compared in terms of accuracy, the Supervised Learning approach and the Dirichlet distribution approach that determines **M**aximum **L**ikelihood **E**stimate (MLE)s, with the latter performing significantly worse than the former. Overall, the validation works as predicted, the run-time of the *Conditional Language* is in the **m**icro **s**econds (ms) range compared to a direct implementation even for several 100,000 data records and the run-time of the validation with about 2000 rules is also fast due to the use of vectorized arithmetic and comparison operations and is less than 7 minutes for 100,000 data records.

# Kurzfassung

FASTA-Daten sind nutzerbezogene Fahrzeugdaten, die zur Verbesserung der Qualität von BMW-Fahrzeugen und zur Weiterentwicklung zukünftiger Motoren verwendet werden. Bislang werden die Daten für Analysen und Auswertungen ohne einheitliche und formale Validierung verwendet. Dies birgt die Gefahr von Fehlern und Doppelarbeit bei Validierungsansätzen. Um diesem Problem entgegenzuwirken, geht es in dieser Thesis darum, die Rohdaten mithilfe des Validierungsframeworks Pandera [1] mit formalen Validierungsregeln auf bekannte Anomalien zu validieren. Um die Erstellung, Erweiterung, Anpassung und Kommunikation der formalen Validierungsregeln zu verbessern, wird eine DSL mit dem Namen *Conditional Language*, entwickelt. Es handelt sich um eine Aussagenlogik DSL, die auf formalen Methoden aufgebaut ist. Sie überprüft eine Pandas [2] Datenstruktur anhand der benutzerdefinierten logischen Sequenz, die auch arithmetische Operationen zulässt.
Die allgemeine Umsetzung basiert auf der Erstellung einer Pipeline, die zunächst die rohen FASTA-Daten lädt, sie dann mit Pandera in Kombination mit der *Conditional Language* validiert und die Daten aufgrund von Verstößen gegen die definierten Validierungsregeln für zukünftige Analysen und Auswertungen kennzeichnet. Um die große Menge an Histogrammdaten zu validieren, werden zwei Klassifizierungsansätze in Bezug auf die Genauigkeit verglichen, der Supervised Learning Ansatz und der Ansatz der Dirichlet-Verteilung der MLEs bestimmt, wobei Letzterer deutlich schlechter abschneidet als Ersterer. Insgesamt funktioniert die Validierung wie angedacht, die Laufzeit der *Conditional Language* liegt im Vergleich zu einer direkten Implementierung selbst für mehrere 100.000 Datensätze im ms-Bereich und die Laufzeit der Validierung mit ca. 2000 Regeln ist aufgrund der Verwendung von vektorisierten Rechen- und Vergleichsoperationen ebenfalls schnell und beträgt weniger als 7 Minuten für 100.000 Datensätze.