# Ensuring Data Integrity and Reliability in Big Data

Metódy inžinierskej práce 2023/2024

Michal Zrutta

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave

26. november 2023

# Motivation

These days data comes too fast and in an enormous sizes, this requires better understanding of data and processing. The upward trajectory poses challenges in data quality. This study delves into the challenges of data quality assurance in the realm of Big Data analytic, exploring methodologies to validate diverse and extensive data-sets. The research focuses on gas consumption in various car models, integrated with AI.

## Outline

1. What is realm Big Data?
2. Problem Statement
3. Quality assessment for big data
4. Methodology
5. Conclusion

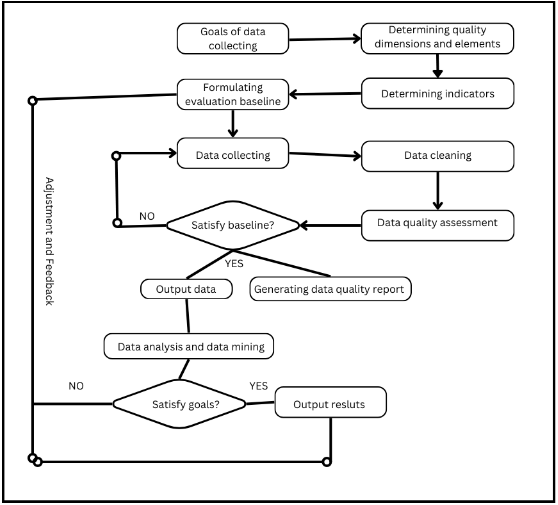
## Realm Big Data

- Massive amount of information
- Sources of data
  - Social media
  - Sensors
  - Online transactions
  - Multimedia
- Uniqueness of Big Data
- Everyday use in industries

# Problem Statement

- Wrong data = *change* in overall results
- How can we ensure data quality?
    - Data Validity
    - Data Completeness/Accuracy
    - Data Consistency

# Quality assessment for big data



Quality assessment for big data. . .

# Methodology

- Literature Review
- Understanding the problem
- Identification of similar approaches
- Innovation through Integration
- Validation through a hierarchical data quality standard

# Data assessment with AI

## Example of gas consumption

- **Defining** data collection goals aligned with strategic objectives
- Implementing a **user ranking system**
- **Cleaning** and **sorting**
- Here plays AI great role
  - AI **automatically** identifies and corrects mistakes
  - AI deep understanding of physics
  - Self-learning mechanism

# Data quality indicators

| Dimensions | Elements | Indicators |
|---|---|---|
| Availability | Accessibility | Data access interface provided, easily accessible for public use or purchase |
| | Timeliness | Data arrival within schedule, regular updates, and timely processing to release |
| Usability | Credibility | Sourced from specialized organizations, regularly audited by experts,<br>and within known/acceptable value range |
| Reliability | Accuracy | Data provided accurately represent the true state of source information,<br>with unambiguous representation |
| | Consistency | Concepts, value domains, and formats remain unchanged after processing |
| | Integrity | Clear format meeting criteria, consistent structural and content integrity |
| | Completeness | Deficiency impact on multi-component use, accuracy, and integrity |
| Relevance | Fitness | Data may not fully match the theme but illuminate one aspect,<br>Most retrieved datasets meet user needs and match their retrieval theme |
| Presentation Quality | Readability | Clear and understandable content, meets user needs, and adheres<br>to specifications in description, classification, and coding |

## Conclusion

- Effectiveness in ensuring data quality
- Reliability and accuracy
- Importance of a systematic approach to data quality assurance