Dr. Mindaugas Šarpis

Lessons on Data Analysis from CERN

Lecture 2

Introduction to Data Analysis

What is data analysis?

Data analysis is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Wikipedia

Key Ideas

- Any experiment (study or analysis) in any field of science will have a data analysis component
- This is the stage of interpretation of the results of the experiment.
- Normally, the results of data analysis appear in scientific publications.

Examples of significance of data analysis in different fields of science and beyond

Biomedicine and Genomics

- Genome Sequencing
- Clinical Trials

Environmental Sciences

- Climate Change Models
- Pollution Monitoring
- Biodiversity Studies

Social Sciences

- Economic Forecasting
- Social Behavior Studies

Astronomy

- Observational Data Analysis
- Gravitational Waves

Engineering

- Predictive Maintenance
- Quality Control
- Structural Health Monitoring

Healthcare

- Epidemiology
- Health Policy

Finance

- Stock Market Analysis
- Risk Management
- Algorithmic Trading

Sports Analytics

- Performance Analysis
- Fan Engagement

Steps of Data Analysis

1. Define the Problem or Research Question

- Formulation
- Experimental Design

2. Collect Data

- Data Acquisition
- Data Collection
- Data Retrieval

3. Clean Data

- Data Selection
- Data Stripping
- Data Skimming
- ...

4. Analyze Data

- Data Exploration
- Data Mining
- Statistical Analysis
- Model Building
- Machine Learning
- Classification (...AI...)

5. Visualize the Data

6. Interpret the Results

- Obtaining Results
- Draw Conclusions from Data

Proprietary Software

Programming Languages

Proprietary tools

- Expensive
- Limited in scope
- Lack compatibility
- Lack flexibility
- Easy to learn / use (GUI)

Programming languages

- Open Source
- Free
- Powerful
- Steep leraning curve (CLI)

Analysis Pipeline (Workflow)