Dr. Mindaugas Šarpis

Lessons on Data Analysis from CERN

Lecture 2

Introduction to Data Analysis

What is Data Analysis?

* What is **Data**? (an interactive exercise)

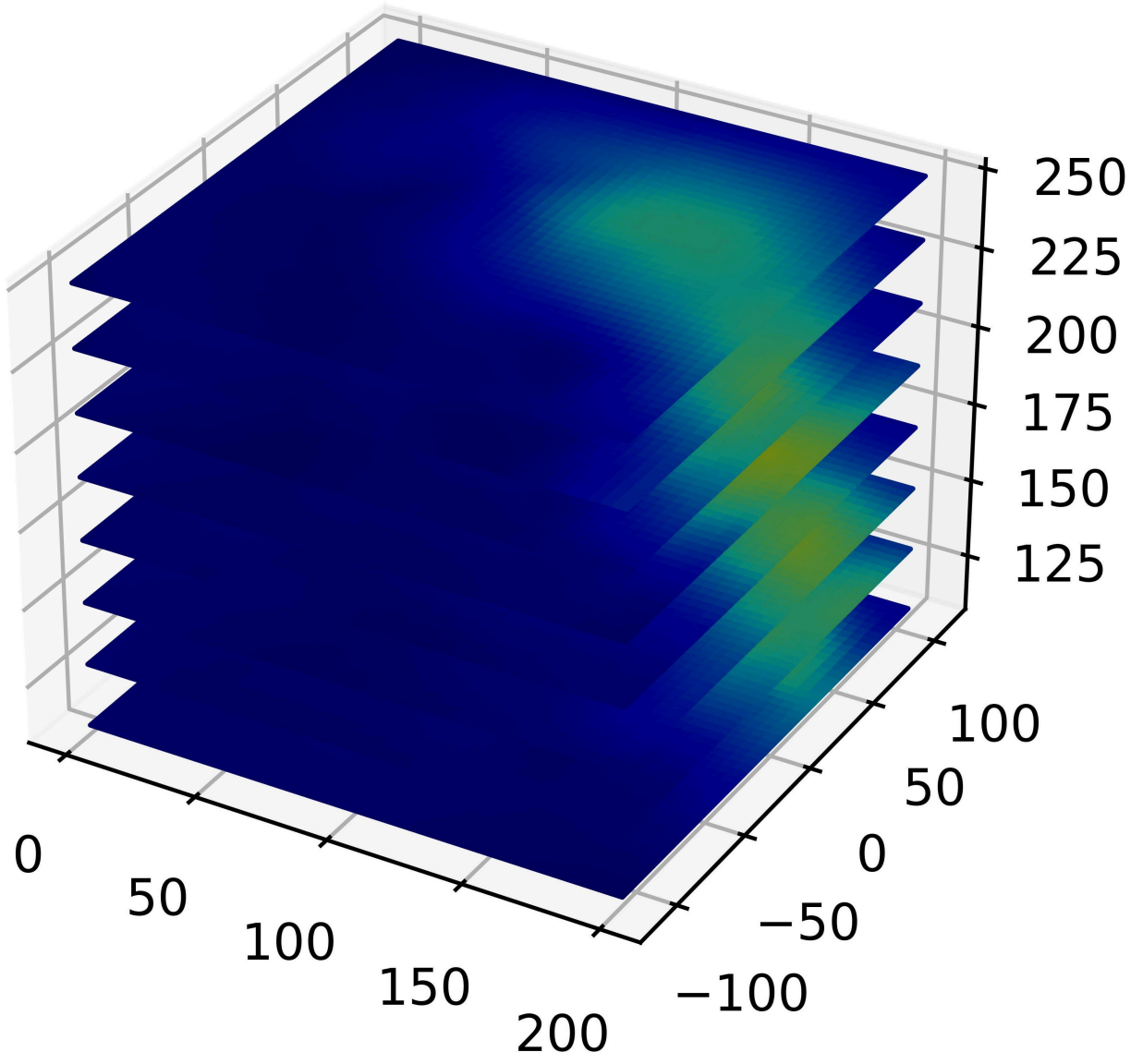
Data analysis is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Wikipedia

What is Data Science?

Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.

Wikipedia


250
225
200
175
150
125
100
50
0
−50
−100
0
50
100
150
200

Key Ideas

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

"*...lacking excellence...*"

"*...aimed at serving the industry...*"

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

Biomedicine and Genomics

- **Genome Sequencing**
- **Clinical Trials**

* 23andMe anyone (ancestry services)?

** comparing against *reference populations*

Environmental Sciences

- **Climate Change Models**
- **Pollution Monitoring**
- **Biodiversity Studies**

* again a *living analysis*

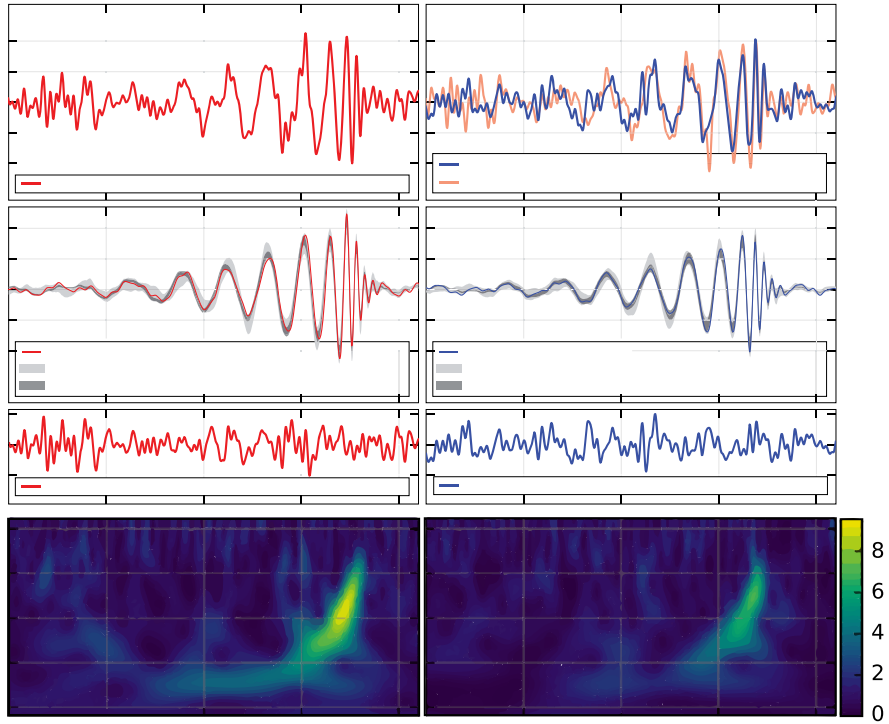
Social Sciences

- **Economic Forecasting**
- **Social Behavior Studies**

* may be qualitative analysis

Astronomy

- Observational Data Analysis
- Gravitational Waves



{ width="80%"

height="auto" }

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**

This might steer the choices in the following steps

2. Collect Data

- How much data do you need?
- What sort of data do you need?
- What data formats should you chose?

3. Clean Data

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

4. Analyze Data

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

5. Visualize the data

- What's your target audience?
- What is the message you want to convey?

6. Interpret and report the results

- Draw Conclusions from Data
- Report Findings

Data Higiene

F A I R

Findable

Accessible

Interoperable

Reusable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

Findable data

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible data

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
    - A1.1 The protocol is open, free, and universally implementable
    - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

Interoperable data

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Reusable data

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
    - R1.1. (Meta)data are released with a clear and accessible data usage license
    - R1.2. (Meta)data are associated with detailed provenance
    - R1.3. (Meta)data meet domain-relevant community standards

Different tools used for data analysis

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)

Discussion

- When to use proprietary tools?
- Whas should you be using?
- saturation of achieved proficiency