DAT121: Introduction to data science

This module makes sure that all students within the 2-year master in data science have a shared acquaintance with certain basics that the higher-level modules can build upon. As an introductory module, DAT121 does not aim at an in-depth discussion or a particularly advanced mathematical presentation of special topics. However, the specific interests and previous knowledge of individual students can be integrated into the module through project work and discussions.

Course days:

- Week 33: Monday (14.8.), Tuesday (15.8.), Thursday (17.8.), Friday (18.8.)
- Week 34: Monday (21.8.), Tuesday (22.8.), Wednesday (23.8.), Thursday (24.8.), Friday (25.8.)
- Week 35: Monday (28.8.), Tuesday (29.8.), Thursday (31.8.), Friday (1.9.)

Schedule for a typical course day:

- 9.15 – 10.00 discussion *(open issues and what was done on the previous day's work)*
- 10.15 – 11.00 lecture – theoretical part
- 11.15 – 12.00 lecture – examples and problem solving
- 13.15 – 15.00 tutorial session *(for project work and solving assigned problems)*

Literature (recommended):

- A. Silberschatz, H. F. Korth, S. Sudarshan, *Database System Concepts*, 7th edn. (international student edn.), McGraw-Hill Education (ISBN 978-1-26008450-4), **2019**.
  (This is the book used for INF230 in autumn block, we occasionally refer to it in DAT121.)
- W. McKinney, *Python for Data Analysis*, 3rd edn., O'Reilly (ISBN 978-1-09810403-0), **2022**. (Open access version: https://wesmckinney.com/book/)
- There is no *required* literature. Any required material will be provided at lecture time.

1. Python basics

The two-year master presupposes basic competency in programming, but not necessarily in Python. Since many data science modules use Python, we begin with a brief **Python intro/refresh**.

Relevant to all modules where Python is used, and specifically to INF201 (autumn semester).

2. Data and objects

We look into techniques from conceptual modelling, such as entity-relationship diagrams, and ontology-based metadata standardization and semantic interoperability. These techniques provide a coherent way to **co-design data and software**, aligning the knowledge graph paradigm (for the data) with the object-oriented or generic programming paradigm (for the software).

Relevant to INF230 (autumn semester), INF205 (spring semester), and INN351 (spring semester).

3. Regression basics

Practical guide to linear regression in Python. This includes parameterizing non-linear models, conducting validation/testing, and giving a basic interpretation to results and their significance.

Relevant to most of data science, especially STAT200 (January block).

4. Good practice (and bad practice)

Introduction to good practices in dealing with data (FAIR, CARE, and RIOT principles) and analysing data. The latter includes a critical discussion of known bad practices and related epistemological issues such as abuse of the *p*-value significance metric, and ideas for how to do better.

Cross-cutting relevance to all of data science.

5. Multidimensionality

Two techniques are introduced for dealing with problems that are multidimensional, but not extremely high-dimensional (*i.e.*, relating three to ten scalar quantities to each other, not thousands or millions): Dimensional analysis, which reformulates a problem in scale-independent dimensionless quantities, and multicriteria optimization, based on the Pareto optimality criterion.

Relevant to many application areas, *e.g.*, IND310 (autumn) and almost all engineering or physics.