Analysis of information systems for AM

BSc Elective Package

Offered by: Department of Mathematics and Computer Science
Language: English
Primarily interesting for: BSc in Applied Mathematics (AM)
Prerequisites: Students are assumed to have basic skills in logic, set theory, calculus, discrete mathematics, databases, programming, and algorithms.


Content, composition, and prior knowledge

Analysis of information, data, and knowledge is increasingly important, with broad application across science, engineering, society, and industry. To tackle these challenges, knowledge and skills in the management, mining, and analysis of (big) data collections is necessary. This elective package provides deeper study of the foundations and applications of analysis of data and information systems.

Students should select a coherent package of courses from the following list. The courses can be followed in any order, but need to take the specific prerequisites for each course into account when scheduling this package. For example, 2ID70 has as prerequisite either 2ID50 in Q2/E or JBI050 in Q2/B.

<table>
<thead>
<tr>
<th>Course code</th>
<th>Course name</th>
<th>Scheduled (Quarter/Slot)</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ID70</td>
<td>Data-intensive systems and applications</td>
<td>Q3 / D</td>
<td>3</td>
</tr>
<tr>
<td>JBI100</td>
<td>Visualization</td>
<td>Q2 / D</td>
<td>2</td>
</tr>
<tr>
<td>2IIG0</td>
<td>Data mining and machine learning</td>
<td>Q2 / C</td>
<td>3</td>
</tr>
<tr>
<td>2IOI0</td>
<td>DBL Process mining</td>
<td>Q3 / C</td>
<td>2</td>
</tr>
</tbody>
</table>

A coherent package is defined as follows: 2ID70, JBI100, and 2IIG0 or 2IOI0.

Course descriptions

Data-intensive systems and applications
This course prepares students to meet the new challenges of contemporary data engineering in which traditional assumptions break, where new data models, query languages and programming interfaces are required. In this course, we study how traditional relational database techniques such as indexing, query planning and optimization, transaction management and self-tuning can be made to work on a massive scale of thousands of machines and petabytes of data. We study models of contemporary data-intensive systems, their efficient engineering, and their practical use. These models include scalable data processing platforms (e.g., MapReduce, Spark) and stream processing engines. We discuss why these models were introduced, their relative advantages and disadvantages, how they are engineered, and how to effectively use them in practice.
Visualization
In the visualization course you will learn about the challenges of visually representing data that comes in a variety of forms. Starting from simple primitive data types like categorical, ordinal, or quantitative data, we will have a look into more complex dataset scenarios including relational data like graphs/networks or hierarchies, multivariate data, text data, or trajectory data that contains an inherent spatio-temporal aspect. In this course you will learn about the data processing, data transformation, data visualization, and finally, the interaction with the visual output. To make a visualization interpretable, readable, and intuitive, we will also have a look at perceptual issues like pre-attentive processing, the visual memory, or Gestalt principles. Moreover, a number of laws or no-goes will be discussed to make the diagrams or visualization techniques more perceptually effective.

Data mining and machine learning
The main focus of this course is on the theoretical foundations of Data Mining and Machine Learning. A secondary focus is on low-level practical aspects (e.g. standard implementations of various models and algorithms). After the course the students will be able to:
- define Data Mining
- define Machine Learning
- define all Machine Learning paradigms: supervised, unsupervised, and reinforcement learning.
- work with data
- preprocess data
- identify which Machine Learning methods are the most suitable for a specific learning paradigm
- derive, implement, and evaluate some of the most widely used methods (listed in the course content) for a specific learning paradigm.
- derive, implement, and evaluate some of the most used deep learning models (listed in the course content) and their learning algorithms
- solve a Data Mining problem using a Machine Learning model.

DBL Process mining
In this DBL, students get a chance to get a first glimpse on process mining. Through a practical case, students will learn the basics of data mining in the context of (business) processes and they build a prediction model for process aspects. In the basic course Data Analytics for Engineers, students have seen the basics of data analytics. In this course, we introduce a new perspective into the mix: The process perspective. Data is not just considered as a static object, but temporal aspects are considered.

In this DBL, we look at prediction problems from a process mining perspective. We use data coming from such process aware system to develop a prediction model to predict the remaining processing time of a case. This prediction model is to be developed using Python or Java or any other programming language using whatever data analysis techniques the students in each group think of.