Advanced Data Science elective package

Advance Data Science	
Offered by	Department of Mathematics and Computer Science
Language	English
Primarily interesting for	All students, but most relevant for students with background in BSc in Data Science (DS)
Prerequisites	Required courses: Students are assumed to have basic skills in logic, set theory, calculus, discrete mathematics, databases, linear algebra, programming, algorithms, and data mining and machine learning. Recommended courses: -
Contact person	dr. Dirk Fahland

Content and composition

Analysis of information, data, and knowledge is increasingly important, with broad application across science, engineering, society, and industry. To tackle these challenges, a wide variety of knowledge and skills in managing, mining, andanalyzing of (big) data collections is necessary. This elective package provides deeper study of the foundations and applications of analysis of data as well as the data-driven construction and analysis of various kinds of models.

Students should select 3 out of the following package of courses. The courses can be followed in any order, but need to takethe specific prerequisites for each course into account when scheduling this package. For example, 2ID70 has as prerequisite JBI050 in Q2/B.

Course code	Course name	Level classification
2IOI0	DBL Process Mining	2.
2ID70	Data-intensive systems and applications	3.
2IX30	Responsible data science	3.
2WS70	Advanced Statistical Models	3.
JBM035	Linear Optimization for Data Science	2.
JBM090	Combinatorial Optimization for Data Science	3.

Course description

DBL Process Mining (2IOI0)
In this DBL, students get a chance to get a first glimpse on process mining. Through a practical prediction case, students will learn the basics of data mining in the context of (business) processes and they build a tool to build prediction models for process aspects.
In prior courses, students have seen various techniques for data analysis and model learning from data. 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 which are causally related, i.e. the application for a loan of 20k Euro leads to a credit check being performed by a bank within 2 hours after the application. In this DBL students will work together in groups of 6 using an advanced scrum approach for managing complex projects and it includes a lecture by an external scrum professional. Part of the work for this course is effective project management, presenting results in a concise manner through a poster and providing predictions for a dataset using any data-analysis tool available to the students.

Advanced Data Science elective package

Data-intensive systems and applications (2ID70)
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.

Responsible data science (2WS70)
The course is focused on studying the problems of fairness, accountability, confidentiality, and transparency (FACT) in data science, and data mining and machine learning in particular. One important challenge to face is that machine learnt models typically are not 100% accurate, i.e. in some ways these models are wrong. Thus, it is important to study how we can make a good use of models that are not perfect, how we can understand the strengths and weaknesses of these models, how we can help a decision maker to trust (or not trust) the model or its particular prediction, and how we can get insights into impact of input features and some inner logic of a predictive model. We need techniques not just to explain the decision of a model, but also to uncover and characterize undesired or even unlawful biases in its performance. Hence, the other important challenge to study is how to formally define such biases, how to uncover and quantify them and how to design machine learning solutions that would enable the so-called fair algorithmic decision making by design. On the other side of the spectrum, there are challenges of privacy and confidentiality. We will study the main principles and techniques that have been researched and employed in data mining for privacy preserving and secure computation to induce models from data and to apply them in real-life scenarios.

Advanced Statistical Models (2IX30)
In many practical situations, different types of data are generated. A data set may contain continuous, discrete, and binary data representing certain properties of objects of interest. The applications we will consider are mostly from life sciences, which implies that we will often study measurements on living entities. Understanding the relationships between variables is often considered important. In the case of medical sciences it may help us understand health and disease. This course will discuss statistical methods and models for estimation of associations and for building nonlinear relationships between variables. The theory is applicable to any area of data science.
Specific topics include generalized linear mixed models, Generalized Estimation Equations, Method of Moments, , model diagnostics, model building, and the use of software.

Linear Optimization for Data Science (JBM035)
Linear optimization is one of the fundamental computational tools in Operations Research, and is used for airline scheduling, production planning, and in many other industrial settings. In fact, it has been called one of the mathematical problems "using up most of the computer time in the world". We will first look at how linear optimization models arise from practical decision problems. Next we will consider the links with linear algebra and geometry. This will lead us to an algorithm, called the simplex method, that may be implemented using techniques from linear algebra. Every linear optimization problem has an associated dual problem, that has an economic interpretation in terms of “shadow prices”, and we will look at these ideas in some detail. Finally, we will consider the case where the decision variables should take integer values, and study a techniques for such problems, namely branch-and-bound.

Combinatorial Optimization for Data Science (JBM090)
Combinatorial optimization consists in finding an optimal solution among a finite set of possible solutions. The main objectives of the course are to understand the complexity of different combinatorial optimization problems, model them with Integer Linear Programming (ILP), and learn different solution approaches to solve them. For some problems (the 'easy' ones) fast solution methods exist while for the other problems (the 'difficult' ones)

finding the optimal solution is time consuming and can only be achieved in practice if the problem is 'small'. For the difficult problems one may have to use `heuristics’, which apply "rules of thumb" to find feasible solutions. Heuristics usually lead to `good’ quality solutions but without any guarantee on their optimality.
In this course we will treat the aspects mentioned above as well as a number of concrete problems and solution methods. Among others, we will treat the easy problems 'shortest path', 'maximal flow' and 'assignment', the difficult problems 'traveling salesman', 'knapsack', and 'vehicle routing', and the methods 'greedy', 'branch and bound', 'dynamic programming', and 'local search'. We will also see how to implement some of these methods in Python.