### Analysis of information systems for CSE

**BSc Elective Package**

**Offered by:** Department of Mathematics and Computer Science  
**Language:** English  
**Primarily interesting for:** BSc in Computer Science and Engineering (CSE)  
**Prerequisites** Students are assumed to have basic skills in logic, set theory, calculus, discrete mathematics, databases, algorithms and programming.  
**Contact person:** dr. George Fletcher, http://www.win.tue.nl/~gfletche/

### 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, 2IX30 has as prerequisite 2IIG0 given in Q2/C and 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>2IX30</td>
<td>Responsible data science</td>
<td>Q3 / C</td>
<td>3</td>
</tr>
<tr>
<td>2I0I0</td>
<td>DBL Process mining</td>
<td>Q3 / C</td>
<td>2</td>
</tr>
</tbody>
</table>

A coherent package is defined as follows: if 2I0I0 is chosen as DBL in major, then package needs to contain 2ID70, JBI100 and 2IX30; else 2ID70, JBI100, and 2I0I0 or 2IX30.

### 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.

September 2019  
Analysis of Information Systems for CSE, BSc elective package
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.

Responsible data science
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.

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.