



## **MSC Information Studies Data Systems Project 2019-20**

### **Project Overview**

The University of Amsterdam Information Studies Master program has a central course that links the practical and theoretical components of the degree around a client orientated live project. The “Data Systems Project” course runs from September - January culminating in presentations at the DSP-20 symposium at UvA Science Park on 29<sup>th</sup> of January 2020.

The course is structured as an extended research period followed by an intensive case study project over a period of 4 weeks. The students work in groups of not more than 5 students and design, implement, and evaluate an interactive system. The application has to fulfil the requirements provided by a client, where the focus lies on the provision of an interactive solution to a big data problem. Students will take roles as project team members and project managers. The working groups will be jointly composed from students of the data science and information systems tracks.

### **Objectives**

The educational objective of the Data Systems Project is to expose students to the practicalities and complexities of real-world projects. Encouraging them to apply the theory, technology and research process skills they have learned to practical problems generated from the needs of real clients and people in real scenarios.

### **Timing and process**

Students are formed into teams of 5 and work through September and October researching a selected problem domain, people, stakeholders and technologies. Through November and December they are expected to engage in early stage prototyping and idea generation with a presentation of proposed designs, functionality and technical requirements before the Christmas break.

In January students work full time on production of practical, working prototypes for testing and evaluation. At the end of January all student teams show their work at the DSP-20 symposium in formal presentations and in demo sessions.

### **Deliverables**

Each student team is expected to create and test a prototype aligned with their client/stakeholder defined project. The teams will each generate a short (3 minutes) video presentation, an academic poster describing their project, and the best 8 teams will provide a formal on-stage presentation at the DSP-20 symposium. After the seminar presentations each team will submit a formal academic report on the research, project development, testing and evaluation.

### **What can be expected by stakeholders**

Whilst it must be clear that finished, deployable products are not the primary objective, and are largely outside the scope of the project, the prototypes produced by student teams contain valuable research and development that can usefully inform subsequent projects or work as early stage investigation prior to commercial development projects. The student teams are creative and innovative with broad skills and high technical awareness but they are often necessarily lacking in experience, the main thing this project helps them learn. It is aimed for that interested students can continue with their work during the thesis project, which runs from April – July 2020.



## MSC Information Studies Data Systems Project 2019-20

### Benefit for stakeholders

Different stakeholders have benefitted from involvement in this project in differing ways. Some stakeholders with clear, immediate needs have received working practical prototype systems that can be tested and deployed onsite, others have received valuable research into audiences, users, technologies or problem domains related to their proposed projects. Other stakeholders have found benefit for themselves and their staff from being involved in an exploration of the needs of their organization and how they formulate this internally and externally.

### Commitment from stakeholders

The time commitment from Project stakeholders can vary though a commitment to certain contact points through the project is expected.

### Key Dates \*

Wednesday 28 <sup>th</sup> August	Initial project presentation
Mid October	Mid term interim presentations
End December	Prototype presentations
Wednesday 29 <sup>st</sup> January	09:00-17:00. DSP student seminar

\*October and December dates due to project organization. Locations at UvA Science Park 904, Amsterdam

### Contact

Dr Frank Nack [nack@uva.nl](mailto:nack@uva.nl)

### Further reference

To see more about the Data Science and Information Systems tracks at UvA DS: <http://studiegids.uva.nl/xmlpages/page/2019-2020/zoek-opleiding/opleiding/5815/233568>

IS: <http://studiegids.uva.nl/xmlpages/page/2019-2020/zoek-opleiding/opleiding/5815/233570>

The Data Systems Project specification:

<http://studiegids.uva.nl/xmlpages/page/2019-2020/zoek-vak/vak/73506>



## MSC Information Studies Data Systems Project 2019-20

### Project description template

#### Project Title

Discrete event simulation for metrics creation

#### Project Stake holder (name and or organization)

ING

#### Stakeholder contact details

Name of organization contact	Pinar Kahraman
Name of organization	ING
Address	Haarlerbergweg 13-15 1101 CH Amsterdam
Tel:	
Mobile:	0657760730
Mail:	pinar.kahraman@ing.com

#### Brief project description

Due to privacy concerns (which increased with the introduction of GDPR), ING keeps the data it creates/collects within its premises. Therefore it is rather costly to store large amounts of machine metrics, curbing data availability, which has been a challenge for IT analytics. The main motivation for this project is to have real-life-like IT metrics to be able to do feed our R&D and analytics initiatives with meaningful data.

The project aims to create a large scale database of annotated machine metrics in the form of time series. This database, if is deemed adequate for analysis, will later be used in a PhD project to test deep learning methods on. The metrics will be an output of a discrete event simulation model of a realistic system with a certain workload, such as log-ins or incoming transactions to be processed, and topology of servers communicating with each other. Independent failures and cascading failures will be simulated in the system to see the effect on the machine metrics, later to be annotated as anomalies.

The teams will work in collaboration (we envision 2 competing teams, built up of 4 squads of 5 students each working together), to handle different aspects of the project, getting together regularly for updates and merges. The set-up of these teams and the aspects to be handled by each team will lie entirely with the students.

The simulation tool, with sample annotated datasets are the main deliverables for this project.

#### Key challenge/problem or message to be communicated within the project domain

The discrete event simulation should not be a very complicated model, as in, the behavior of the model should be thoroughly understood, but should still be realistically representative of a real computer system that handles workload of the nature that ING handles.



## MSC Information Studies Data Systems Project 2019-20

### **Exemplar projects or additional reference materials (books, papers, products, URLs etc)**

The students will need knowledge about the logic and logistics behind workload processing utilized by computer systems (a good book is: “Principles of Transaction Processing” by Philip A. Bernstein, Eric Newcomer), model building, parameter sensitivity analysis, and discrete event simulation. As an example of a stochastic simulation tool, students can take a look at <http://jmt.sourceforge.net/>.

### **Goals or key criteria for an explorative project**

Understanding the logic and logistics behind workload processing utilized by computer systems is very important in the explorative phase.

### **Suggested key requirements or success factors for an implementation**

Ideally, the simulation will be able to handle different types of workload that have different types of processing logic underneath, which result in different types of behavior. It is also of interest to have the input load volumes follow a real-life-like pattern specific to the type of workload (*e.g.* that would entail daily and weekly seasonality). Random incidents will be introduced here and there, which should be annotated in the output time series data. Of course, the type of incidents may also differ, which in turn will again produce different types of responses.

The simulation should spit out time series data, as well as the annotations when the incidents occur, and we would like to be able to produce and store large volumes of it. Therefore, the simulation tool should be fast enough, and the user should be able to introduce different parameters to produce time series with different behavior.

The students can decide how they want to handle the user interface of their tool, where potential modules and configurations can be set by the user. And most importantly, results should realistically reflect machine behavior. Therefore, a comparative study with real datasets would be very welcome.

### **Challenges or constraints envisioned (if any)**

### **Any specific technical or content requirements**

### **Any additional comments or thoughts**

We envision working together with different teams that have different objectives to be the most challenging part of this project. However, students should bear in mind that this is also the case in companies, where dependencies and communication generally form the most challenging parts of project management. Therefore, we believe that it will be an excellent chance for them to learn about how to handle this aspect of business and discover where their personal talents and challenges lie within collaborations.