DATA SCIENCE

Masters Programme Focus Area

Department of Industrial and Systems Engineering

Johan W. Joubert
johan.joubert@up.ac.za
T +27 12 420 2843

Elias J. Willemse
elias.willemse@up.ac.za
T +27 12 420 3443

Our modern world is filled with data, from personal information like our shopping habits, to data generated continuously by technology, such as smartphones, that are fully integrated with our lives. It is estimated that 70% of the data currently available, did not even exist 4 years ago.

Organisations are faced with a new challenge of how to extract value from all the available data sources and make more informed decisions that will improve their business operations, and maintain their competitive edge. This need has given rise to a new discipline.

Data Science is the competency to make sense of, and find useful patterns within data to better support decision-making. It is a natural extension and honing of those quantitative skills that have always been part of an Industrial Engineer's armour. The bar is raising to impact the organisation.

At the University of Pretoria, we provide a structured, well-defined path for students to complete an Industrial Engineering Masters degree, focussing on mastering the skills of Data Science, and applying it to a cutting-edge research project.

Students enrolling for an Honours degree with the aim to continue to Masters can further supplement the programme through carefully selected Honours modules.

Master the following

1. **Formulate context-relevant questions**
The keyword in data science is not data; it is science. You want to start with a question that is worth asking. Does it add value? Is it pitched correctly to support decision-making?

2. **Get and handle large data sets**
Big Data abound. The challenge is rather in how you go about sorting through the variety of sources, from the web, APIs, crowd-sourced data, and cleaning it up into a tidy, workable set that is relevant to your question.

3. **Exploratory data analysis**
Before jumping in and building a suite of models, the first step is to slice and dice the data in insightful ways so that you can get a thorough understanding of what the data does, and does not allow.

4. **Develop rigorous decision models**
In the majority of decisions we make there is both variability and uncertainty. Learn the skills to develop scientifically rigorous models that are reproducible, and offer reliable results.

5. **Data visualisation and interpretation**
Too often we are just too glad to have been able to produce results. The value, however, is in presenting the results (or just the original data) in an insightful, intuitive way.

6. **Packaging data products**
Tell compelling stories from your data and results. Automate complex analysis, or just apply technology to share data with larger audiences.

Tuks of niks
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Masters Programme Focus Area

Available, scoped projects for 2014/5

1

Making sense of, and finding useful patterns within data is at the heart of data science. In this project, the objective is to extract driving behaviour profiles from large accelerometer data sets. The goal is to cluster context-specific variables and vehicular events and characteristics to identify not merely plausible, but evidence-based profiles that are useful for decision-making such as customer profiling, driver incentive schemes, and real-time driver behaviour feedback.

2

In a co-evolutionary setting, the state of the art agent-based transport simulator, MATSim, allows for a rich description of behaviour. One subpopulation we model is that of logistics carriers. Carriers are allowed choice dimensions over which they autonomously can try and optimise their daily plan, for example: which customers are assigned to which vehicles; routing of those vehicles; and fleet sizing. Contrary to traditional Vehicle Routing Problems (VRPs), which are solved on static networks, this project is scoped within the dynamic multi-agent setting where carriers have to solve these tricky problems amidst other decision-makers who too are trying to optimise their routes and fleets. Instead of developing a VRP algorithm, you will evaluate the impact of using different variable and fixed cost models using existing algorithms. How do different costing models affect the resulting fleet composition? The candidate will learn valuable skills in data science that deals with making rigorous decisions amidst the uncertainty and variability of solutions. These skills are valuable not only for this particular project, but will provide insight to the candidate in developing robust solutions in future.

3

One of the state of art, agent-based models used for transport planning, namely MATSim, requires a sophisticated synthetic population. Each "agent" in the population represents a commercial vehicle, and the vehicle's activity chain for the day is derived from complex network theory. That is, something similar to social network theory's six-degrees-of-separation. We often overlook the importance of the statistical sampling methods that we use. This masters dissertation is about studying the consequences of using two different sampling regimes when generating a synthetic population from a given weighted, directed complex network. The successful masters candidate will gain insight into the managerial and decision-making implications of just changing the quantitative assumptions. How do we ensure, with scientific rigour, that decisions we make are both reliable and reproducible?

4

As a logistics service provider, the point of collection or point of delivery often accounts for much wasted time. In this dissertation the objective is to study the spatiotemporal characteristics of activities, and identify potential bottleneck facilities. This is achieved by exploring the manoeuvring of vehicles through their activity chains. Past research have extracted detailed activity chains for more than 40,000 commercial vehicles over a six-month period. For this project, the candidate will study the activity chains and consider the location, and frequency of (potentially) wasteful activities. The masters candidate will be exposed to a number of data science elements. Not only will you be dealing with large data sets, but you will be exposed to, and required to visualise it in useful ways to answer managerial questions.

5

Residential and commercial waste collection is a basic, yet costly service provided by local governments. In South Africa, metros alone spend in excess of R4-billion annually on waste collection. The objective of this dissertation will be to develop a waste generation model, that is, estimating how much waste is generated where. The skills you will develop in this project is to programatically (and periodically) extract data from various sources such as GPS records, crowd-sourced land use data, and point values from disposal site weigh bridges, and extracting useful patterns from integrating those data sets. The final goal is to develop a data "product" that will provide the inputs for short-term operational decisions such as routing, medium-term decisions like sectoring of residential areas into service-days, and long-term decisions like the (re)location of landfill sites.

6

To understand logistics operations across multiple and diverse company types, we can employ Business Intelligence (BI) to transform raw event data into meaningful information. Events are merely the result of underlying behaviour and, deeper still, the structure of the logistics system. In this dissertation the objective is to poke at a large (given) data set of commercial vehicle activity chains, and evaluate the similarity of the chains over multiple days. The masters candidate will be exposed to a number of data science elements that are not just necessary for the completion of the project, but that are valuable in any evidence-based decision-making contexts in future. These include dealing with large data sets; formulating context-relevant questions that are both valuable from a scientific and managerial point of view; exploratory data analysis; generating reproducible results; and packaging the data in a way that allows for useful managerial decision-making.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi