D3M PhD research proposal 2020

Supervisory team

Prof.dr. Didier Fouarge, ROA, d.fouarge@maastrichtuniversity.nl
Dr. Rui Jorge de Almeida, KE, DAD, BISS, rj.almeida@maastrichtuniversity.nl
Dr. Niels Holtrop, MSCM, n.holtrop@maastrichtuniversity.nl

Research project

Title: Machine Learning approaches to labour demand and educational choices

Relevance and impact: The demand for skills is changing rapidly due to technological disruptions. The rapid development in information and communication technology took a big turn, particularly from the 1970s and 1980s onwards, when more and more routine production tasks could be pre-programmed and carried out by machines or by cheap - often foreign - low-skilled workers, who became more easily deployable as a result of increasing globalization (Autor et al., 2003). This had major consequences for the labour market position of medium-skilled workers, who previously carried out such routine production tasks and suddenly became redundant (Goos & Manning, 2007) but also affects the demand for non-routine analytical skills and interpersonal skills at the higher end of the labour market (Weinberger, 2016; Deming, 2017). Consequently, there is evidence that the returns to education increasingly vary by field of study (Altonji et al. 2016). Youngsters who choose for the 'wrong' field might end up unemployed or in low paid jobs with little security (flexible employment/platform jobs). This PhD project aims at further advancing Machine Learning techniques for a better understanding of the skill structure of labour demand, well-informed educational decision making and causal inference. It does so in three related research projects using data on the demand for skills (from online job vacancies), the supply of skills (from descriptions of higher education programmes), and study/occupational choice (form a large career orientation private company).

Project 1 focusses on information extraction from vacancy data from Dutch websites (Jobfeed). Based on the state of the art literature, we will identify skills associated to routine and non-routine tasks in intermediate level (mbo) and high level (hbo+) jobs. We use text mining (e.g., Hartmann et al., 2019), named entity recognition (e.g., Lample et al., 2016) and information extraction (e.g., Sarawagi, 2008) to systematically code demanded skills in online vacancies in order to quantify the demand for routine and non-routine skills, as well as possible relationships between these skills across occupations. This allows us to 1) characterize occupations w.r.t. their skill content, 2) quantify the diversity of skill domains in occupations and 3) characterize the skill similarity between occupations. While questions (1) and (2) were addressed earlier, the only academic effort to quantify the task similarity between occupations (3) we are aware of is the survey-based task similarity indicator by Gathmann and Schönberg (2010). The impact of this project is that we inform the market and job seekers (e.g., through unemployment office UWV) of different occupations with related skill content. This information is relevant to avoid occupations with declining demand and/or declining pay and conversely, in identifying occupations with growing demands and/or high pay.

Project 2 focusses on using Machine Learning to guide study choice. Choosing a study programme is a complex process. Economic theory states that students choose the study that maximizes expected utility, which depends on their skills, preferences and labour market opportunities. Psychologists, however, argue that young people are less rational when choices are complex and often have incorrect expectations about study contents, their level of ambition, cognitive requirements, and salary. Career orientation and guidance (LOB) programmes aim to support young people to make the right study and occupational choice. This project questions how well LOB programmes support young people in making 'good' study choices, i.e., with low probability of drop-out and/or switching to a different field (when the match with preferences/abilities is poor), and fields associated with high labour demand.

As part of an ongoing project, Fouarge works with Qompas, the largest LOB programme in preparatory vocational education (vmbo) in the Netherlands. The programme consists of 13 modules with questions on experiences, occupational preferences and personality amongst others. Not all schools use all 13 modules, and they are not implemented at the same time in all schools. This year, the student data in Qompas will be matched to DUO data so that the educational track and field of study choice in vocational education (mbo) can be measured at the individual level. We will use Machine Learning (ML) models to predict indicators of drop-out/switch and identify 1) school characteristics associated with which LOB modules are implemented at what time (these characteristics come from the Qompas system, but also include socio-economic characteristics of the school neighbourhood), 2) individual characteristics and LOB modules that are strongly associated with the field of study choice in mbo, and 3) key individual characteristics, LOB modules and their timing that are associated with poor choices (high drop-out/switch, or choosing for fields that are associated with occupations with little demand as identified in project 1), and good choices (switches that are high in demand and/or lead to high paying occupations with similar skill descriptions as identified in project 1). This model will subsequently be used together with Qompas for a field study using a new cohort of students, to change the LOB modules and their timing in a RCT framework to assess the causal effect education choices, evaluated using causal ML techniques (e.g., Wager and Athey, 2018). The impact of this project is that we develop early warning indicators that inform those young people who are about to make study choices that are of poor quality either because they are associated with high drop-out, switch rates or poor labour outcomes (unemployment, flex work, platform work etc.). At the same time, we provide assurance to those young people who make good quality study choices, reducing the stress that more than 50% of young people experience due to major decisions such as study choice (Stemmingmakers, 2018).

Project 3: Non-routine skills in higher education: a Machine Learning approach. Automation changes the demand for skills. Routine skills become less important, as these skills can easily be performed by computer technology. However, non-routine skills such as analytical, problem solving, creative and 'people' skills grow in importance. This project questions how the skills offered in university programmes match to the skills (and combinations of skills) demanded by the employers on the labour market (project 1).

We answer this question in several steps. First, we gather programme descriptions of university programmes in the Netherlands. These descriptions can be obtained from, e.g., QANU, VSNU and university websites (e.g., using web-scraping techniques). We use text mining (e.g., Hartmann et al., 2019), named entity recognition (e.g., Lample et al., 2016) and information extraction (e.g., Sarawagi 2008) to quantify the extent to which non-routine skills play a more or less important place in university programmes (at the CROHO level), i.e., the supply side of skills. Second, we use the information from this first step and combine it to information about the skills demand for highly educated employees obtained from project 1 to assess the expected rate of labour market success of graduates by CROHO field. Third, using administrative data from statistics Netherlands for all university graduates who graduate in year t, we monitor early labour market indicators in t+1 and t+2 (e.g., likelihood of job, wage, job mobility) and relate labour market success to the extent to which their education programme offered skills that match the demand for skills. One would expect that, other things being equal, graduates for CROHO codes that are richer in non-routine skills would perform better. The impact of this project is that we develop predictive indicators of the extent to which the supply of skills in higher education matches the demand.

Contribution to literature: The use of Machine Learning in the field of labour and education economics is still very limited. For example, to date at IZA-Institute of Labor Economics, the largest worldwide network of labour economists, only counts six publications out of 13,285 with "Machine Learning" in title or abstract. We contribute to this thin literature, which will ensure a high visibility for our output. From a technical point of view, this project will make advances in information extraction to identify phrases that describe skills sets, occupations and their relations

in job advertisements. Furthermore, this project will also focus on matching textual information from different sources in the labour market, which are possibly defined using different terms and phrases.

Societal value: This project 1) will inform the market (e.g., UWV) about the similarity in the skill content of jobs, 2) improve the way occupational guidance is provided (Qompas), and 3) will enable us to derive guidelines for designing university programmes that fit the skills' demand on the market (e.g., useful to universities, VSNU). This work will also advance the use of Machine Learning in supporting field studies as well as improving decision in a data driven manner.

Link to D3M: The members of the supervisory team are part of the D3M research community. We contribute new knowledge to areas relevant to D3M such as workforce analytics, empirical micro- and labour economics, scaled consumer analytics for decision makers and Digital platforms. We will share our findings with and gain feedback from the D3M community during seminar presentations. This project will also tie the D3M community to the Learning & Work community for which the findings of this project are highly relevant, as they will show how to implement Machine Learning methods in the fields of labour and education economics.

Multidisciplinary perspective: Fouarge is a labour economist who specializes on changes in the demand for skills in the labour market and how this affects occupational and educational choices. Almeida is a specialist in Machine Learning and Text Analytics, focusing on developing data driven models from structured and unstructured data to assess hidden patterns and complex dynamics. Holtrop is an applied econometrician with a specialization in consumer choice, unstructured data (text and visual), and Machine Learning for causal inference. Fouarge and Holtrop currently work together on a research project that uses survey data and text analytics to identify skills supplied in study programmes of universities of applied sciences in the Netherlands. The lessons learned from this ongoing cooperation will be of use for a swift progress in project 3. Almeida is working on several academic and industry Text Analytics and Natural Language Processing projects, which will have cross-fertilizations with this project. Holtrop has applied Text Analytics in several academic projects, and teaches this material in the BISS course Unstructured Data Analysis. Our project also involves the cooperation with a private company (Qompas).

Planned studies and timeline

Scientific publications: 5 scientific publications in international peer-reviewed journals: 1 publication for project 1, 2 for project 2 and 2 for project 3. For both projects 2 and 3 we aim at 1 technical publication in the field of Machine Learning that focusses on methodological advancements, and 1 publication in an journal in the field of economics that focusses on the substantive findings.

Presentations: 4 presentations to the D3M community to which we convey interested researchers from Learning & Work. 8 presentations at international conferences. We will target conferences in the field of business analytics as well in labour and education economics. We foresee 6 presentations for stakeholders (e.g., Ministries of OCW and SZW, UWV, Vereniging van Schooldecanen en Loopbaanbegeleiders) to share our findings.

Professional publications: We will reach out to stakeholders in non-technical publications for professionals in the field, e.g., ESB, Didactief, DecaZIne, Tijdschrift voor Hoger Onderwijs, Qompas Nieuwsbrief.

Timeline and cooperation: The cooperation is as in the table below that shows that success requires joint expertise from three involved departments. Given the complexity and multi-method (ML and experimental) nature of project 2, more time is reserved for this project.

	Project 1	Project 2		Project 3	
		Technical paper	Substantive paper	Technical paper	Substantive paper
Fouarge	X		X		X
Almeida	X	X	X	X	
Holtrop	X	X		X	X
Timing	Year 1/2	Year 1/4		Year 2/3	

Funding

The GSBE funding only includes limited budget for data acquisition. For access to CBS/DUO data (project 2 and 3), we ask the PhD candidate to write a funding proposal as part of the Odissei programme (https://odissei-data.nl/en/en-odissei/). To cover costs of data specialists at Qompas, and conform to the wish of SBE to acquire additional funding, PhD and supervisors will submit a grant proposal to OCW (directie Kennis). If additional resources are needed, the PhD candidate will submit a research proposal to ROA, which has reserves to invest in promising data collection.

Five key publications of the supervisory team

Almeida, R. J., Baştürk, N., Kaymak, U., & Sousa, J. M. (2014). Estimation of flexible fuzzy GARCH models for conditional density estimation. *Information Sciences*, 267, 252-266.

Fialho, A. S., Vieira, S. M., Kaymak, U., Almeida, R. J., Cismondi, F., Reti, S. R., ... & Sousa, J. M. (2016). Mortality prediction of septic shock patients using probabilistic fuzzy systems. *Applied Soft Computing*, *42*, 194-203.

Fouarge, D., Kriechel, B., & Dohmen, T. (2014). Occupational sorting of school graduates: The role of economic preferences. *Journal of Economic Behavior & Organization*, 106, 335-351.

de Grip, A., Fouarge, D., Montizaan, R., & Schreurs, B. (2020). Train to retain: Training opportunities, positive reciprocity, and expected retirement age. *Journal of Vocational Behavior*, 103332.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing*, *34*(1), 154-172.

Profile of the PhD

The PhD candidate has strong computer skills and is trained in data analytics. The candidates has strong affinity in applying ML skills in the field of labour and education. For being succesful, the PhD requires supervision and training in ML, labor economics and choice models.

References

Altonji, J. G., Arcidiacono, P., & Maurel, A. (2016). *The analysis of field choice in college and graduate school: Determinants and wage effects*. In Handbook of the Economics of Education (Vol. 5, pp. 305-396). Elsevier.

Autor, D., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118(4),* 1279-1333.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, *132*(4), 1593-1640.

Goos, M., Manning, A., & Salomons, A. (2014). Explaining job polarization: routine-biased technological change and offshoring. *American Economic Review, 104*(8), 2509-2526.

Gathmann, C., & Schönberg, U. (2010). How general is human capital? A task-based approach. *Journal of Labor Economics*, 28(1), 1-49.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20-38.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, *1*(3), 261-377.

Stemmingmakers (2018), "Hoe kies jij je studie?", available at https://stemmingmakers.nl/index.php/2018/10/03/studiekeuze123-en-stemmingmakers-onderzoeken-studiekeuzestress/

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

Weinberger, C. (2014). The increasing complementarity between cognitive and social skills. Review of *Economics and Statistics, 96*(4), 849-861.