This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Creating a typology of analytics Master's degrees in UK universities: Implications for employers and educators

PLEASE CITE THE PUBLISHED VERSION

<https://doi.org/10.1080/01605682.2019.1605468>

PUBLISHER

Taylor & Francis © Operational Research Society

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the Operational Research Society on 14 June 2019, available online:
<http://www.tandfonline.com/10.1080/01605682.2019.1605468>.

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Mortenson, Michael J., Neil Doherty, and Stewart Robinson. 2019. "Creating a Typology of Analytics Master's Degrees in UK Universities: Implications for Employers and Educators". Loughborough University.
<https://hdl.handle.net/2134/37533>.

Creating a Typology of Analytics Masters Degrees in UK Universities: Implications for Employers and Educators

Michael J Mortenson

WMG, University of Warwick, Room 111, Engineering Management Building,
Coventry, West Midlands, CV4 8UW, UNITED KINGDOM
m.mortenson@warwick.ac.uk

Neil F Doherty

School of Business and Economics, Loughborough University, Loughborough, LE11 3TU, UNITED
KINGDOM
n.f.doherty@lboro.ac.uk

Stewart Robinson

School of Business and Economics, Loughborough University, Loughborough, LE11 3TU, UNITED
KINGDOM
s.l.robinson@lboro.ac.uk

In recent years there has been a growth in specialised analytics Masters degrees, in the UK and beyond. However, there has been little research into the contents of such degrees. In particular, the role disciplines such as operational research play within them remains an under-explored area. Using a mixed-methods approach, this paper analyses UK Masters degrees in analytics to determine a typology of provisions. Firstly, a support vector classifier is used to identify the traditional disciplines analytics degrees most closely align with. Secondly, a hybrid approach to analyse the modules included in analytics curricula is employed, as part of which a new metric (module topic weighting) is presented. The analysis identifies two main categories of degrees, the first aligning with machine learning and computing topics; the second operational research and business themes. The paper concludes by evaluating the implications this has for students, employers, educators and the operational research discipline.

Keywords: analytics, support vector machines, curricula development, machine learning

1. Introduction

The rapid explosion of data generated in the modern world, “2.5 quintillion bytes of data every day” (Marr 2018), has been matched by a growing demand for staff equipped to manage and analyse this data (e.g. e-Skills and SAS, 2013; Manyika *et al*, 2010; Rae, 2018). Meanwhile, researchers associated with a variety of academic disciplines, such as information systems (e.g. Chiang *et al*, 2012) and, of particular relevance to this study, operational research (e.g. Liberatore and Luo, 2010), have pointed to the synergies between their discipline and analytics, and the potential of their degree courses to meet this demand. Further, many universities have sought to meet this need through specialised degrees, with titles such as “Business Analytics” or “Data Science”, which have proliferated throughout Europe, North America and the rest of the world. A recent report from Deloitte identifies over 100 analytics-related degree programs in the US alone (Danson *et al*, 2016).

Indeed, the value of these ventures has been acknowledged by many governments of such countries. In the forward to a UK Government report, David Willetts and Matthew Hancock (then Ministers for Universities & Science, and Skills & Enterprise respectively) state:

[The] potential impact [of Big Data] is so significant that it could transform every business sector and every scientific discipline. [...] The challenge of meeting the demand for skilled people, from both industry and academia, is one that is globally recognised. It is a challenge that cannot be tackled by government in isolation, which is why we will work with industry and academia to come up with solutions.

HM Government (2013)

This discussion demonstrates the potential importance of analytics courses (alongside other training activities), to meeting the perceived skills gap associated with big data and analytics. Less clear, however, is the content that such courses should contain and their overlaps with other existing disciplines. This issue is the central theme of this research and is explored by quantitatively and qualitatively analysing the course materials associated with Masters degrees in analytics and related areas from UK universities.

More specifically, the objectives of this research are to:

- RO1.** To identify the academic disciplines with which analytics Masters degrees in the UK most closely align;
- RO2.** To determine the specific skills, subjects and techniques taught within analytics degrees;
- RO3.** To evaluate the current academic response to analytics; the implications this has for employers (and graduates seeking employment) and educators, and also for the operational research discipline.

To do this we present a relatively novel approach to analysing degree materials through a large-scale quantitative exploratory analyses of all (as of time of writing) the Masters degrees in analytics in the UK (and related disciplines such as statistics and OR). Based on our literature search, we find no directly comparable approach has been employed, and in general, despite the suggested importance of such research, this is deemed to be a relatively under-explored area.

The choice of limiting to UK universities only has obvious drawbacks. In the modern university-landscape, and at post-graduate level particularly, this is increasingly an international marketplace. For analytical reasons, evaluating course contents in multiple languages is significantly more complex. However, English-language courses in analytics are obviously available outside of the UK; in Australia, Canada, the US, as well as in countries where English is not the main language. However, whilst a limitation, there are several reasons for this design choice. Firstly, the process of collecting course materials is a manual one (detailed in the next section), and therefore very time consuming. Secondly, this research is part of a larger project specifically investigating analytics in UK universities. Finally, despite the internationalisation of higher education, there is likely to be local differences, so expanding to include all English-language degrees may introduce extra complexity, bias and noise.

In meeting these objectives we believe a clear contribution will be made on three fronts. Firstly, the work can contribute to a clearer understanding, in the UK higher education system at least, of the academic response to the need for analytics graduates – an understanding that can help shape future course design in this and in other areas. Secondly, by providing some practical help to employers and potential students in identifying the most suitable courses (albeit, with the caveat that the results can only truly speak to a UK market). Finally, as this effectively represents a comparison between analytics and other disciplines, or at least one facet of this (academic training), the results can help to define better this growing area of interest and position it alongside related areas of enquiry, most notably operational research (OR).

To this end, the remainder of this paper is arranged as follows. In the second section we evaluate the discussion of analytics education in the literature. Thirdly we present the methodology used in this study. The fourth section reports the results of the first analysis conducted, a classification of analytics degrees to the discipline they most closely align with (thus meeting RO1). The fifth presents an analysis of the modules within analytics degrees, the skills and techniques associated with these, and, again, links these to traditional academic disciplines (meeting RO2). The sixth section presents a typology of analytics degrees and details the core differences between these. The final section considers the practical, methodological and theoretical implications of the study (RO3), and the contributions that are made, before finally some concluding remarks are given, with discussion of limitations and future research.

2. Analytics and Big Data: The academic response

A review of the current literature regarding analytics degree curricula suggests three major themes. Firstly there are a variety of papers, as alluded to in the introduction, which seek to identify the disciplines that inform and overlap with analytics. Secondly, there are specific examples given of how a course can, or should, be constructed, and the topics and techniques that should be covered. Finally, there has been research analysing existing degree curricula, work that has the same intentions as this paper. Each of these will be discussed in sequence.

2.1 Analytics and Related Disciplines

For the reasons given in the introduction, many academic disciplines stand to gain significant traction from an association with analytics. Unsurprisingly, therefore, many have sought to highlight synergies between analytics and their field of work, as well as the potential value this may offer, including in the OR community. Liberatore and Luo (2010, p 313) asks how OR “practitioners can take advantage of the surging interest in analytics to promote the OR profession and expand its reach”. In relation to the information systems (IS) discipline, Chiang *et al.* (2012, p 6) argue analytics represents “a unique opportunity for [IS] units in business schools to position themselves as a viable option for educating professionals with the necessary depth and academic rigor to tackle the increased complexity of [analytics]”.

A variety of authors have suggested various “taxonomies” of related disciplines (e.g. Varshney and Mojsilovic, 2011; Chiang *et al.*, 2012; Evans, 2012). For the purpose of this study, we will refer to the version given in Mortenson *et al.* (2015) and shown in figure 1. This representation divides analytics disciplines into three areas *technologies*, *quantitative methods* and *decision making*; that is analytics in effect relies upon a combination of skills in these three areas such that disciplines that feed into each of these may have relevance for analytics graduates and analytics degree courses.

Figure 1 A Taxonomy of Disciplines Related to Analytics (Source: Mortenson et al, 2015)

From this perspective, analytics can be viewed as the intersection of a variety of pre-existing disciplines, and therefore, to some extent, we can view analytics degrees as being the combination of different elements of each discipline. This view is expressed in Evans (2012, p 6):

As academics in business schools, we have been teaching these topics for over 40 years, albeit in a disjointed and compartmentalized fashion. Business analytics provides the framework to exploit the synergies between traditionally-diverse topics in a more practical, application driven format.

Of course, inasmuch as this is an opportunity, it is also a challenge; determining which elements of which disciplines should be included and how the teaching can be combined into the limited timeframes of a degree program is not an insignificant task. Proposals to meet this challenge that have been presented in this literature will be the focus of the remainder of this section.

2.2 The Creation of Analytics Programs

Creating degree programs to meet the needs of analytics employers is a challenge that can be addressed in three ways:

1. Through modifying degrees in related disciplines to also incorporate some of the techniques, use cases and contexts of analytics. In other words, these would be traditional discipline degrees, albeit somewhat adapted to address some of the concerns of analytics employers and to incorporate some explanation of the role of that discipline in analytics practice.
2. Through creating specialisations within programs so to offer some proportion of 'analytics' content, as well as that of a traditional discipline. There are several examples of this approach, such as the University of Texas' *MBA: Business Analytics Concentration*.

3. Through creating bespoke “analytics” degrees, examples of which will be presented later in this paper.

In respect to the first of these, Chiang *et al.* (2012) discuss the potential for IS courses to evolve to meet this demand, and the extent to which this requires curricula to be developed, but also the focus to be changed. They identify a list of (primarily quantitative) modules that would need to be included, as well as a requirement to focus on “rapid interpretation and business decision making based on huge volumes of information” as opposed to an orientation towards “the management of transaction data and the production of information for management” (Chiang *et al.* 2012, p 5).

The second approach, one which is closely related, is to provide specialisations or a collection of ‘elective’-type modules, again within existing disciplines. An example of this is presented in Molluzzo *et al.* (2015), whereby the authors create a list of modules to be included in such a concentration targeted at IS students. However, a caveat to such efforts, and those towards modifying existing degree courses, is given in Chen *et al.* (2012), again in reference to the IS discipline. Ultimately graduates of such schemes are more likely to still find roles within IS groups, albeit with a greater awareness of analytics and big data and its uses in organisations.

The alternative, of providing bespoke courses that combine aspects of the different disciplines involved in analytics, is less explicitly discussed in the literature. In part this absence may be due to the political nature of such a task; ultimately this infers a ‘shared ownership’ between discipline groups. In other words, this is a trans-disciplinary task, and not necessarily one that suits the specialised nature of academic journals and research, therefore potentially limiting the opportunity to publish such work. However, it is a task seemingly important to understanding analytics courses, and to determining the proportions of each of the related disciplines that should be covered.

2.3 The Content of Analytics Curricula

The aforementioned study into the design of a ‘data science’ specialisation for IS students (Molluzzo *et al.*, 2015) seeks to identify the modules that should be incorporated in such a program. Their research was based upon identifying recurrent topics over “the online syllabi of 21 introductory courses that contained Data Analytics or Data Science in their titles” (Molluzzo *et al.*, 2015, p 13). Notwithstanding the contribution of their work, there are two important differences between their study and ours. Firstly their area of focus is on data science specifically and on introductory courses; this research will seek to evaluate courses that have titles associated with terms such as “analytics”, “big data” and “data science” and on the basis of full (graduate-level) degrees in the area rather than modules. Secondly our focus is on linking areas of analytics-orientated study to pre-existing disciplines; that is to better determine the root elements from which analytics degrees draw upon.

A second example of work in this area, is a series of surveys “to assess academia’s response to the growing market need for students with Business Intelligence (BI) and Business Analytics (BA) skill sets” conducted by the Association for Information Systems (Wixom *et al*, 2014, p 1). In this research a large scale questionnaire ($n = 1,379$) was conducted of university staff, students and practitioners. One area of exploration in the study surrounds the modules that respondents (those involved in educational provisions) currently offer at their university. Again this presents some interesting results, however, such modules could be featured in a wide range of different degrees, and no indication is given as to whether these are electives or core units; what academic level they are at; or how such courses combine to offer the full range of required skills and understanding required of analytics graduates. Secondly, due to the survey instrument, there is a possible concern that different respondents may interpret the different categories differently, and therefore introduce some subjectivity into these findings.

2.4 Summary

In summary, the literature has demonstrated that this is an important area of study, and that the issue is problematised by the involvement of multiple disciplines in the analytics field, and the subsequent political issues and competition between fields this entails. However, to date, there appears to have been no significant empirical research into bespoke analytics courses and their contents, the traditions from which these modules are drawn, and, by implication, the relative importance of different disciplines upon the analytics curricula. This is a gap this research seeks to fill, and the methodology applied to do so is presented in the following section.

3. Methodology and Data Collection

Analysing degree curricula materials is a task that can be performed in many ways. Firstly, a purely algorithmic approach can be employed, although no examples were found of this in respect to curricula data in the current literature. Whilst this can fail to identify some patterns that are recognisable to the human eye, equally there are other structural patterns that may be more easily found in such an approach. Additionally these methods can eliminate many of the judgement calls that qualitative methods necessitate, decisions that can lead to potential biases in the results. Finally, a benefit of such an approach is the ability to ‘scale up’. In qualitative analysis the time taken is almost directly linear to the number of documents. For algorithmic approaches most of the time overhead is in the setting up and validation of the model, such that these approaches mean a far greater quantity of materials can be analysed whilst remaining within a reasonable time frame.

Secondly, ‘hybrid’ methods can be employed where (qualitatively derived) dictionaries of codes are built and applied to degree curricula to identify (quantitative) patterns (e.g. Chu, 2006). This

approach has obvious benefits, particularly as it allows the researcher to ‘correct’ the data by finding synonyms and patterns that are far harder to identify algorithmically, due to the complexities of text data as a whole. For example, a common approach is to create taxonomies of associated terms such that the researcher can record word counts in a topic rather than the words individually. A principal benefit is that, for a knowledgeable reader, the words “masters”, “MS” and “MSc” can be regarded as the same, but for a machine this similarity is harder to identify without prior instruction. However, equally there are drawbacks. Whereas in much of the prior research the authors have smaller scope, and therefore can build taxonomies purely on *a priori* theory, as alluded to in the introduction, we conceive analytics to be a wide-ranging practice encompassing many different traditional disciplines. Similarly these disciplines are not perfectly partitioned. So, for example, the keyword “data” has linkages with IS, computer science, statistics and OR (among many others). In such case we either need to disregard the word entirely or to include it as a count in each of the individual categories. Whilst this can seemingly resolve our problem, the relative counts of a word such as “data” could be very significant in identifying the properties of each discipline.

Finally, and most commonly in the literature, traditional qualitative methods such as ethnography, cases studies or textual analysis can be used to evaluate the contents of course materials and the teaching methods used (e.g. Stern, 1998). Such methods allow for a deep investigation of the topic, but, as trade-off, the scale of the investigation is likely to be limited (due to time resources) and there is the potential for subjective bias to be introduced.

As this discussion highlights, each approach has its merits, but equally they have their drawbacks. In light of this, our methodology sought to apply all three (to some extent). We firstly performed a large-scale quantitative analysis of the data using machine learning techniques (an algorithmic approach) to meet RO1. Thereafter, we performed a closer, ‘hybrid’ analysis of the results by creating coded ‘themes’ to detail the modules the degrees offer (RO2). In both cases we used qualitative content analyses to validate and explain the results. Each of these approaches are discussed in sequence, before finally the data collection and transformation processes are detailed.

3.1 A Large-Scale Quantitative Analysis of Analytics Degrees

The first (purely algorithmic) analysis is designed to meet the first objective of this research (RO1), identifying the academic disciplines to which existing analytics degrees most closely align. In respect to the scale of the task, and the debate above, this was conceived as a classification problem.

There are many different methods that have been used to algorithmically classify text documents, but some of the most common include random forest ensembles (Ho, 1995), k-nearest neighbour

classification (e.g. Fukunaga and Narendra, 1975) support vector classifiers (Cortes and Vapnik, 1995), and logistic regression (Cox, 1958).

However, as neatly summarised in the principle of the “no free lunch theorem” (Wolpert and Macready, 1997), in practice there is no way to know which algorithm will work best for such a task “up front”. Accordingly, the four methods mentioned are all used in conjunction, as part of an ensemble classifier where each has a vote on class assignment and the final decision is made by committee. There are two main approaches to such voting, hard and soft. In the case of the former, a single, discrete vote is given by each algorithm, whereas in the latter a probability is given. As all the algorithms can support soft voting, this is the approach favoured in this research. However, the performance of each algorithm is not expected to be equal. Accordingly, after the first run of the algorithm, the accuracy scores of each were recorded, and applied as a weighting to the voting process. That is, an algorithm that is 70% accurate is weighted by 0.7 in terms of voting; while an algorithm with 40% accuracy is weighted as 0.4.

The overall approach, therefore, was to use a range of classifiers, based on each of the listed algorithms, and use these to predict a class label for course materials associated with Masters degrees titled “analytics”, “data science” or similar (e.g. “business analytics” or “Big Data analytics”). In other words, we are seeking to classify analytics-type degrees on the basis of their similarity to traditional academic disciplines, the results of which can indicate the relative prominence of the teaching of these disciplines. The class labels were drawn from a range of disciplines related to analytics using the taxonomy of disciplines shown in figure 1. Whilst this is by no means a complete list of related disciplines, it was felt this would give a broad cross section of the core areas involved. Therefore, the list of disciplines was determined as:

- **Computer Science (CS):** Technologies
- **Information Systems (IS):** Technologies / Decision Making intersection
- **Psychology (PS):** Decision Making
- **OR:** Decision Making / Quantitative Methods intersection
- **Statistics (ST):** Quantitative Methods
- **Machine Learning (ML):** Quantitative Methods / Technologies intersection

An additional issue with each of the algorithms listed is the selection of certain free hyperparameters, the choice of which can significantly impact the accuracy of the model. These parameters were optimised using grid search (e.g. LeCun *et al*, 1998), whereby different values are tested on different folds of the data (using 4-fold cross-validation), and the values which produce the highest averaged accuracy are retained. The hyperparameters optimised for each, and the values tested (with the optimal value indicated with an asterisk), were as follows:

- **Random Forest:** *Number of estimators (tested for 50, 100* and 150).*
- **K-Nearest Neighbours (kNN):** *Number of neighbours used (tested for 3, 5 and 7*).*
- **Support Vector Classifier (SVC):** *The value of C, the penalty parameter (tested for 0.5, 1.0* and 1.5), and for kernel (linear*, radial basis function and polynomial)*
- **Logistic Regression:** *The value of C (tested for 0.5, 1.0 and 1.5*).*

Our final concern is data quantity. For this study we chose to limit the course materials to Masters degrees offered at UK universities. This was partially due, as discussed in the introduction, to reducing the possible variation that may come from national and/or language difference between materials from universities from different countries, and partially due to the time-consuming nature of collecting data of this kind. However, the accuracy of machine learning predictions tends to increase as the size of the training dataset increases, and with only 234 course descriptions the size of the data set would be considered relatively small.

Our solution is to seek data that can be used as a 'proxy' for course descriptions; documents that can be classified in the same way (by the disciplines above) and which display comparable characteristics as degree materials in respect to the similarities and differences observed between categories.

A likely candidate for such criteria is job adverts related to the respective disciplines. It could be assumed that both OR degree materials and job adverts will include terms such as "optimization" and "simulation" at a far higher frequency than those associated with information systems, and, vice versa. They are less likely to feature terms such as "data warehouse" or "ERP". Whilst, intuitively, job adverts and degree descriptions would read very differently, this should have little impact on their usefulness for this task. Essentially the central mechanism of the model is based upon the terms which most distinguish the different categories, not the terms that distinguish the documents. In other words, it is the relative use of terms such as "optimisation" and "ERP" that will distinguish class membership, not the relative use of terms such as "salary" and "lectures".

Of course the implicit assumption is that the Masters degrees linked to each discipline are actually aligned to the jobs they are intended to prepare students for, an assumption that some may well question. Therefore, in order to validate job descriptions as a suitable surrogate for the course materials, the model, trained on job advert data, can be tested against a subset of the job advert data as well as the course materials. If the accuracy of the model does not significantly decrease when using the course data to test the model, the job advert data can be taken to represent an appropriate proxy.

With such data, there are significant risks in terms of data quality. This is not only due to the data source (essentially internet data), but also the fact that we are using data from two different sources

(as above), and also that the classes themselves are not as clearly separated as they would be in most classification tasks. In other words, the difference between a machine learning course/job and a computer science one, is not as clearly delineated as would be the case in most classification tasks (as, arguably, machine learning is a subset of computer science). Similarly, disciplines such as computer science and information systems, or statistics and OR, will have significant overlap.

Accordingly, an additional measure to deal with what is effectively outlier control was included. As a prior step to classification, both the job advert data, and the degree material data associated with six academic disciplines listed above, was first analysed using a clustering algorithm, HDBSCAN (Campello *et al*, 2013). HDBSCAN has several properties that make it attractive to our use case. Firstly, like DBSCAN (Ester *et al*, 1996) on which its based, it allows the discovery of density clusters, even within other clusters. This is particularly useful when considering the interconnectedness of some of the source material. However, HDBSCAN also includes a hierarchical component, in the style of hierarchical clustering techniques such as Ward's method (Ward, 1963), which allows the algorithm to find clusters of lesser density.

The algorithm was first set to process job advert data, with classes (e.g. job advert discipline search terms – “CS”, “IS”, “ML”, “OR”, “PS”, or “ST”) provided to the algorithm. For this reason, considering the sparsity of text data as a whole, the provision of class labels was expected to (and was able to) assist the clustering algorithm towards finding the “right” classes.

However, as a clustering algorithm that is somewhere between “hard” (partition based) and “soft” (data can be between multiple clusters) methods, the approach allows us to effectively generate a probability of class assignment. The inverse of this probability is effectively the chance that the item (job advert) is an outlier. However, in comparison to “normal” outlier analysis, in this case we can establish outliers at not only a global level (across the whole dataset), but also a local level (within clusters). For example, we can identify a “CS” job advert that is not an outlier in respect to the whole dataset, but is an outlier compared to other “CS” job adverts. After scoring every job advert using this method, we excluded any that scored in the bottom quartile in terms of assignment probability (i.e. keeping the top 75% in probability of assignment). Such an approach means that we only retain the job adverts that are closest to ‘normal’ in regards to their classification.

Having completed this process for job advert data, we can use the model to then predict course materials. Through doing so we eliminate the bottom quartile of discipline course materials, in respect to their ‘outlier-ness’ for each cluster assignment (based on the job advert data). In other words, we reduce the overall data by 25%, removing the items that are furthest from cluster assignments (which are obviously influenced by the declared discipline class).

The full process for this research method, therefore, is as follows:

1. Fit a clustering algorithm on the job advert data, with labels provided, so that the bottom 25% of records, in terms of probability of assignment, are removed;
2. Use the clustering algorithm to predict course materials for each class, again removing the bottom 25% of results;
3. Select a random subset of both job advert data and degree materials for training and for testing;
4. Optimise the hyperparameters for each algorithm using grid search;
5. Ascertain the accuracy of each algorithm and weight its vote based on this percentage and build the final model;
6. Use the model to predict the remaining subset of job advert data and course materials (the test set) and evaluate its accuracy;
7. Use the model to predict the classes for analytics-type degrees and evaluate the results.

The procedures described were performed in Python, using *HDBSCAN* (McInnes *et al*, 2016), *scikit-learn* (Pedregosa *et al*, 2011) and *NumPy* (van der Walt *et al*, 2011), and the *Python Data Analysis Library* (PyData, 2012).

3.2 Weighted Module Analysis

The second (hybrid) approach, designed to meet RO2 (identifying the skills, subjects and techniques taught within analytics degrees) combines both qualitative and quantitative elements. Firstly, the modules incorporated in each program were extracted from the degree materials. Secondly, the modules drawn from the disciplines associated with analytics (i.e. all the degrees except those labelled as “analytics”) were coded based on their principal themes. The codes were created inductively, and iteratively updated during the progress through the corpus. The counts of the occurrence of each code (one per module) were retained. This allows a comparison of the influence of the different disciplines in the dataset of analytics degrees, by identifying the module topics most closely linked to each discipline and computing an “association score” for each.

One concern in this process is the level of granularity to apply in this coding. For example, a module in *C# programming* could be coded as specifically this title, as *C programming* (i.e. the whole family of coding languages), *programming* as a whole, or even a more generic code such as *computing*. In general the approach used was to seek a reasonably granular level of detail in the codes deemed to be most important to analytics as a whole (drawing on the existing literature on analytics), so we chose the code label *C programming* in the specific example given here, and then broader codes for the topics that are less directly related (for instance, both “brand marketing” and “international marketing” were labelled as “marketing”). The second main concern was around modules that covered topics which incorporated more than one code. For example, during the analysis separate codes for *machine learning* and *data mining* were created. In order to classify a module titled “machine learning and data mining”, we simply allocated a count of 0.5 to each code.

Although some insight can be gained from simply analysing the overall counts, many topics will be recurrent across multiple disciplines, and also due to the different quantity of materials extracted in respect to each discipline (shown in table 1), counts alone may not represent the relative frequency across disciplines. Therefore we sought to introduce a scoring system that takes into account the relative importance of each code (module topic) in comparison to other disciplines; in other words the codes which had the most discriminatory power in characterising each of the degree types.

To do so, borrowing from the widely used χ^2 test, we compute an 'expected' count for each term in each discipline; that is a calculation of the frequency one would expect if the term was distributed proportionally across the different disciplines. We can then compare the actual frequency of the topic in each discipline to its expected count, such that if the actual exceeds the expected, we conclude the topic is important to the discipline. However, we also need to control for the fact that some disciplines have a greater number of topic codes (primarily due to the disparity in sizes of datasets). Therefore we finally take the amount that the actual exceeds the expected as a proportion of the total quantity of observed occurrences to give what we describe as a *Module Topic Weighting (MTW)* for each term and each discipline. This approach can be written algebraically, with *MOF* as the Module Observed Frequency and *MEF* as the Module Expected Frequency, as:

$$(1) MTW = \frac{MOF - MEF}{MOF}$$

Where (with *TMD* as Total Modules in Discipline; *TFT* as Total Frequency of Topic across all disciplines; and *TM* as Total Modules):

$$(2) MEF = \frac{TMD}{TM} \cdot \frac{TFT}{TM} \cdot TM$$

There are two important differences between these equations and those used in the χ^2 test. Firstly, in contrast to equation (1), in the χ^2 test one would square the difference between observed and expected and divide by the expected. In this instance we do not need to square the difference (as we are only interested in positive results) and we measure as a proportion of the observed rather than the expected as we are more interested in the frequency of the term than in the scale of the difference. Secondly, we do not seek to compute a test statistic to compare to the χ^2 distribution. In essence this is an omission that is forced upon us. The common rule of thumb, that expected counts should be at minimum of five, would be violated in the vast majority of cases due to the fact that the number of variables (topics) exceeds the number of cases (degrees). Whether this represents a major change in respect to statistical validity is open to debate, but ultimately we treat the results of this analysis as an indicator rather than statement of fact, and represents a part of a series of methods upon which we reach our conclusions. However, to ensure that the figures are not

distorted by outliers, we only retain *MTW* scores for terms that occur more than twice in each discipline, and represent greater than 1% of the total topics of that discipline.

Having created *MTW* scores for each topic in each module list of the degrees associated with the disciplines listed (but excluding analytics), this now gives a quantitative basis on which we can assess the relative importance of different degrees in the modules offered in our analytics degrees. To do so we applied the same coding structure to the module lists of the analytics degrees and again retained the counts. We then multiply the frequency of that topic across analytics degrees by the *MTW* score associated with each discipline and finally summing these by discipline. As such, we ultimately produce a final score for each discipline (which is a combination of the frequency of the topics in analytics degree modules and the relative importance of these topics to degrees in that discipline) from which we can compare the relative influence of the discipline on analytics degrees.

3.3 Data Collection

As discussed, the data source used is degree materials. These were collected by manually checking each UK university's website directly (using the Guardian's University League Table¹ to determine the institutions to include) and extracting written materials on the relevant Masters-level courses. In cases where universities offered more than one degree with a related title then only the one deemed most relevant was used (e.g. "Business Psychology" was deemed more relevant than "Clinical Psychology"). In cases where 'hybrid' degrees were offered, such as an MSc in "Operational Research and Applied Statistics", the result was excluded on the basis this would complicate classification. Additionally, we also collected materials from degrees related to analytics, including titles such as "Business Analytics", "Big Data" and "Data Science". In this instance multiple degrees for the same institution as well as 'hybrid' degrees were retained, as this information was considered useful to our goal, and also the results were qualitatively assessed after the analysis. As discussed above, for the ensemble model (RO1) the course material data was supplemented with job adverts linked to each discipline which is used as a proxy due to the relatively small size of dataset. The job adverts were extracted using the job search of the social network LinkedIn (www.linkedin.com), via its application programming interface, with each of the disciplines listed above used as the respective keywords for the searches. 1,500 job adverts were randomly selected from the keyword searches made for each discipline. The total quantity of all documents extracted is shown in table 1.

Table 1 Quantities of Documents Collected by Discipline and Type

¹ <http://www.theguardian.com/education/ng-interactive/2014/jun/02/university-league-tables-2015-the-complete-list>

Discipline	Code	Job Adverts*	Degree Materials
Computer Science	CS	1,477	69
Information Systems	IS	1,494	40
Machine Learning	ML	1,415	16
Operational Research	OR	1,481	10
Psychology	PS	1,500	37
Statistics	ST	1,479	19
Analytics	AL	---†	43
TOTAL		8,846	234

* The quantity of job advert data varies due to the removal of duplications

† Analytics job advert data was not extracted as this is not used as part of the model build, and solely as the 'validation' set

An additional, but critical step when dealing with text data (particularly when sourced from the internet), is the performance of data cleansing and transformations. These included removal of duplicates from the job advert data, as well as stopwords (very frequent words with little discriminatory power), numbers, punctuation and all HTML/XML code. The remaining words were stemmed (whereby each term is reduced to its shortest root such that “business” and “businessman” are both reduced to “busi”). Finally, the dataset was transformed into a vector of word counts by document (a Document-Term Matrix) and weighted using the term frequency-inverse document frequency (TF-IDF) algorithm (Salton, 1991) so that each word is weighted by its relative importance in the document. Full details of these steps are presented in the appendix.

4. Analysing Analytics Degree Curricula

The two forms of analysis described in the previous section were performed in sequence, with this section detailing the results of the first, the classification of analytics degrees to related ‘traditional’ disciplines (meeting RO1). As per the steps listed in the methodology, the first stage was to perform HDBSCAN clustering on the job advert and course materials data, and filter out the worse performing quartile in terms of probability of cluster membership (local outliers) as well as any that are not assigned to any cluster (global outliers). Having done this, the volumes per class were as shown in table 2.

Table 2 Quantities of Documents by Discipline and Type (Post Outlier Removal)

Discipline	Code	Job Adverts	Degree Materials
Computer Science	CS	1,052	58
Information Systems	IS	918	36
Machine Learning	ML	1,055	13
Operational Research	OR	1,065	9
Psychology	PS	1,500	16
Statistics	ST	1,045	11
Analytics	AL	----	43
TOTAL		6,635	186

It is worth some discussion of the impact of outlier removal on the dataset (table 2). As this is performed across the whole data, as opposed to on a class-by-class basis, data has not been removed in equal proportions. Psychology remains untouched, with all 1,500 retained, whereas Information Systems, the most impacted, is reduced from 1,494 to 918 (removing 38.5%). The other classes have reduced by around 30%. The implication, and one which makes some theoretical sense, is that Psychology job adverts are more distinct from the other classes.

This result does have implications, and leads to some concerns. Firstly, the result shows some between-class imbalance, which can have negative impacts on the later classification model. While Psychology is now over-balanced, at a rate of around 150% of the other classes, this is not so high as to suggest major problems. Indeed, in a literature review of the area, He and Garcia (2009) argue for a definition of imbalanced as instances where the imbalance is “significant”. While no definitions are given as to what “significant” is in this instance, examples are given of ratios of 100:1 or greater. Additionally, because our overall sample size, and individual class-size, remains relatively high, the impacts can be expected to be relatively small.

Secondly, the results of the DBSCAN algorithm may suggest that the Psychology class is more of a fringe case, and less central to our analysis. However, it is important to recall that the data used at this stage is used for training, with the degree materials used in the actual predictions (and to inform this study). It does not necessarily follow that the course materials would overlap in the same way. In fact, a good argument can be made that most of the job advert data could be assumed to have more inter-class overlap (“outlier-ness”) than the course materials, as employers may favour more versatility and interdisciplinarity than may be expected in a master’s degree in a specialised subject. Overall, these results suggest that there may be reason to have some caution as to the relevancy of the Psychology class, but without any guarantee that the same “outlier-ness” would occur in the course materials data, and as the choice of disciplines is to reflect the concepts of earlier work (figure 1), we will maintain this class for now.

The next step was to optimise the hyperparameters (the results of which were detailed in section 3.1) and to assign the appropriate weights to each algorithm based on their accuracy. The overall accuracy of the ensemble, prior to weightings, was 65%, with the following individual accuracy scores:

1. Random Forest: 0.74 (+/- 0.01);
2. k-NN: 0.43 (+/- 0.01);
3. SVC: 0.54 (+/- 0.02);
4. Logistic Regression: 0.65 (+/- 0.01).

These accuracy scores can be used as weightings (e.g. the Random Forest algorithm was weighted as 0.74), an approach known as weighted soft voting. As demonstrated in Zhou (2012, pp 74-77), the weighted soft voting approach is mathematically superior to non-weighted approaches (albeit, assuming that the classifiers are independent, which in practice is unlikely).

In doing so, the overall accuracy of the ensemble is increased to 69%. Whilst this is perhaps less than ideal, it is worth noting that accuracy in multi-class prediction is typically lower than in the more common binary-class prediction. Held in contrast to a completely random, unbiased classifier, which would on average correctly assign the class at a level of $1 / \text{the number of classes}$, our classification level of 0.69 outperforms the random figure (0.167) by a ratio of over 4 to 1. It is also worth noting that this figure of 0.167 is also key in evaluating the accuracy of individual classifiers. There would be little value in including a classifier whose accuracy is worse than, or approximated the random figure. In this instance the worst performing classifier (k-NN) maintains a ratio of 2.58 to 1 which, while not ideal, suggests that it does find something in the data which may support performance. Considering that we are effectively penalising this algorithm via the weighting system, on balance it is deemed worthy of inclusion in the ensemble.

In consideration of this discussion, the classifier was regarded as appropriate to use for our purposes, particularly as its results are further validated with the additional research methods described in the previous section.

Accordingly, we used the algorithm to predict the test data, a mix of both job advert and course materials again. In this instance the prediction accuracy of the algorithm was actually improved, at 75% overall (a relatively healthy figure considering the data and task). The overall results are summarised in the confusion matrix of table 3.

Table 3 Confusion Matrix for the Test Data

		Predicted Class						Total
		CS	IS	ML	OR	PS	ST	
Actual Class	CS	170	52	20	4	7	11	264
	IS	20	70	1	4	8	5	108
	ML	30	3	156	10	12	11	222
	OR	10	13	23	142	22	18	228
	PS	4	15	6	20	612	13	670
	ST	9	8	18	31	16	128	210
Total		243	161	224	211	677	186	

Whilst the overall accuracy is relatively high (75%), analysis of table 3 suggests that the better-than-average accuracy of Psychology classifications has bearing on this. If we were to ignore all Psychology data, and also those misclassified as Psychology, prediction accuracy drops to 69%. This adds, to some extent, to the earlier concerns that Psychology may be somewhat on the periphery of the dataset. Again, however, the more important part remains the application to course material data, which can be explored in more detail.

To further explore performance we calculated further metrics. Three of the most commonly used are (where TP = True Positives; FP = False Positives; FN = False Negatives):

$$\text{Precision} = TP \div (TP + FP); \text{Recall} = TP \div (TP + FN); F1 = 2TP \div (2TP + FP + FN)$$

In other words, precision gives a measure of the proportion of correct predictions out of all of the predictions made for that class; recall gives the proportion of correct predictions by the total number of instances of the class in the test data; and F1 gives the harmonic mean of these two metrics. The results of these metrics are shown in table 4. As can be seen, there is some discrepancy between the effectiveness of the classifier for different disciplines. Psychology is the best performing with an F1 score of around 0.9, whereas information systems scores the lowest at 0.52. When evaluated alongside table 4, there are some instances where “OR” data has been misclassified as “statistics”, and vice versa, as well as similar misclassifications between “computer science” and either “information systems” or “machine learning”. Whilst such misclassifications are understandable, considering the overlaps in these areas, they are problematic. However, in many ways it is more important to consider the performance in respect to just the course material data, as the job advert data was used moreover as a proxy.

Table 4 Precision, Recall and F1 Measures for the Test Data

	Precision	Recall	F1 Score
CS	0.644	0.700	0.671
IS	0.648	0.435	0.520
ML	0.703	0.696	0.700
OR	0.623	0.673	0.647
PS	0.913	0.904	0.909
ST	0.610	0.688	0.646

Predicting, using the ensemble model, just the subset of course materials data, shows significant improvement, with accuracy rising to 97.2%. This equates to only four misclassifications across the 143 course materials included. This is illustrated in the confusion matrix (table 5) and the precision, recall and F1 measures (table 6).

Table 5 Confusion Matrix for Course Materials

		Predicted Class						Total
		CS	IS	ML	OR	PS	ST	
Actual Class	CS	57	0	1	0	0	0	58
	IS	0	36	0	0	0	0	36
	ML	2	0	11	0	0	0	13
	OR	0	0	0	8	1	0	9
	PS	0	0	0	0	16	0	16
	ST	0	0	0	0	0	11	11
Total		59	36	12	8	17	11	

Table 6 Precision, Recall and F1 Measures for Course Materials

	Precision	Recall	F1 Score
CS	0.983	0.966	0.974
IS	1.000	1.000	1.000
ML	0.846	0.917	0.880
OR	0.889	1.000	0.941
PS	1.000	0.941	0.970
ST	1.000	1.000	1.000

Analysis of tables 5 and 6 show that there are still some small concerns regarding the algorithm separating “machine learning” and “computer science” (with two of the former misclassified, and one of the latter), but in general the results are very positive. While the extent of this improvement (97% accuracy here compared to 75% across the whole data) is significant, it is perhaps unsurprising. As noted earlier, it makes some theoretical sense that degree materials may be more closely aligned with the core of their discipline, while job adverts demand a more diverse set of skills, as well as carry more “noise” from an analytical perspective.

When considering the Psychology class, again performance is strong (an F1 score of 0.97), but with the overall increases in performance across the board, these results seem less of an outlier (indeed Psychology would technically be the 4th best performer in this subset). In consideration of these final results, based on the data used to determine our findings, the case for removing this category seems much weaker. The potential downside to its inclusion would be either that it represents noise that limits prediction accuracy, or conversely that it inflates our accuracy statistics as an outlier class. Given the overall accuracy on the subset (97%), and that its performance seems in line with other classes, neither problem seems particularly relevant.

The final stage is then to predict the ‘analytics’ degree materials (thus meeting RO1). Obviously, with no analytics class, there are no accuracy or performance metrics to report, however, we are, of course, interested in the model’s class predictions. Of the 43 degrees, 17 were classified as machine learning; 15 as OR; 5 as information systems; 4 as statistics; and one each for computer science and psychology.

The degree classed as computer science is the University of Essex’s *Big Data and Text Analytics*. Whilst the description clearly incorporates significant analytics/data science elements, there is an emphasis on computer systems (i.e. the internet) as the generator of big data and text data in particular. The degree classed as psychology is Swansea University’s *Management (Analytics)* MSc. From the course description, and indeed even the degree title, there is a clear emphasis on management issues, which likely explains the classification.

The four degrees classified as statistics are Birmingham City University’s *Data Analytics and Management*; Brighton University’s *Data Analytics*; the University of East London’s *Data Science*; and Swansea University’s *Finance and Business Analytics*. One obvious connection between the titles is an emphasis on data (although that is not necessarily surprising in the area), which obviously is also a topic core to statistics. In summary, while there is not a huge amount that is immediately obvious in terms of diagnosing this classification, there is a seeming emphasis on data mining and data analysis in general.

With five degrees classified in this way, information systems is the third most assigned, just above statistics. The degrees included here are the University of Aberdeen’s *Data Analysis, Visualisation and Communication*; Coventry University’s *Data Science and Computational Intelligence*; Manchester Metropolitan University’s *Business Technology and Analytics*; University of Sheffield’s *Data Science*; and University of Westminster’s *Business Intelligence & Analytics*. While there is diversity in these titles, the inclusion of “technology” in one, and “computational intelligence” / “business intelligence” in two others, gives some acknowledgement to systems that would be considered part

of analytics and data science, but at the same time are also obviously relevant to the information systems discipline.

OR is the second highest at 15 classifications. Again, clear patterns can be seen amongst those classed as such. Firstly, there are several degrees that from the title alone would suggest an OR flavour. Examples include the University of Leeds' *Business Analytics and Decision Sciences*, the University of Manchester's *Business Analytics: Operational Research and Risk Analysis* and the University of Lancaster's *Management Science and Marketing Analytics*. There is also more of a "business" orientation in the degree titles, with 10 featuring the keywords "business analytics", and others with related "business" keywords, such as Loughborough University's *Business Analytics Consulting* and Aston University's *Business and Marketing Analytics*. There is an emphasis on business schools as the "host" of these programs. Ten of the degrees are based in business schools, 3 in computing schools (one jointly hosted between business and computing schools), with 1 hosted in mathematical sciences.

The final, and most populace classification is machine learning, with 17 degrees classed in this way (compared to the 15 classed as OR). Where those classed as "OR" frequently incorporated the keywords "business analytics" in their titles, for this classification "data science" is the most common (8 out of 17), followed by "big data" (5 out of 17). Overall, 12 out of the 17 had one of these two keywords in their title. Similarly, while the "OR" classed degrees were mostly hosted in business schools (11 had this association), for this classification 14 of the 17 were hosted in computing/technology schools, one jointly in a technology school and business school, and only two were hosted solely in a business school. Overall there is seemingly more of an emphasis on computational approaches to analytics.

This analysis presented in this section is designed to meet RO1: *to identify the academic disciplines which analytics Masters degrees most closely align with*. However, in doing so the results suggest there are multiple 'types' of analytics degrees, each of which align to different disciplines. While each degree received at least one classification, the majority, by a substantial margin, were classed as one of two categories; either on the basis of a similarity to the OR or machine learning disciplines. Consequently, for the module analysis, which is presented in the next section, we elected to analyse not only the full analytics dataset, but also on the subsets that have been labelled 'OR' and 'ML' in the ensemble model.

5. A Curricula Analysis Using Module Topic Weighting

As detailed in the methodology, the second analysis performed was solely on the module titles contained in the degree materials (designed to meet RO2). Thematic codes were developed and

assigned to each module in the traditional discipline set (i.e. all materials other than those associated with “analytics”). In total there were 106 codes identified across nearly 2,000 modules. Using the *MTW* weighting scheme (presented in the methodology) each discipline’s codes were given a score that sought to capture the discriminatory value of that topic to the discipline in comparison to the others. Using this scheme we therefore identified the most important topics by discipline, as presented in figure 2.

Figure 2 Top 10 Highest Weighted Topics (*MTW*) by Discipline (in descending order of discriminatory power)

Computer Science		Information Systems		Machine Learning	
Term	MTW	Term	MTW	Term	MTW
Graphics	0.5406	Strategy	0.8140	Robotics	0.8416
Distributed computing	0.5224	Performance management	0.7821	Natural language processing	0.8130
Computer architecture	0.5197	Enterprise resource planning	0.7804	Image processing	0.8045
Mobile	0.5057	Management	0.7764	Machine learning	0.8026
Internet programming	0.4947	Information systems	0.7415	Computer vision	0.7691
Software	0.4649	Knowledge management	0.7094	Visualisation	0.7581
Computer security	0.4592	Project management	0.6897	Artificial intelligence	0.7104
Programming	0.4531	Business intelligence	0.6632	Business intelligence	0.6627
Multimedia	0.4304	Operations management	0.6486	Agents	0.6607
Networks & servers	0.4249	Human resources	0.6486	Neuro science	0.5777

Operational Research		Psychology*		Statistics	
Term	MTW	Term	MTW	Term	MTW
Spreadsheets	0.9420	Psychology	0.8954	Bayesian statistics	0.8701
Supply chain management	0.9171	Business psychology	0.8954	Hierarchical data	0.8701
Operational research	0.9137	Social psychology	0.8954	Experiments	0.8701
Decision sciences	0.9033	Cognitive psychology	0.8954	Surveys & sampling	0.8701
Operations management	0.9033	Clinical psychology	0.8954	Linear models	0.8701
Game theory	0.9033	Neuro science	0.8431	Regression	0.8701
Optimisation	0.8943	Human resources	0.7211	Survival analysis	0.8701
Consulting	0.8937			Geospatial	0.8701
Simulation	0.8922			Monte Carlo	0.8701
Stochastic modelling	0.6713			Medical & health	0.8588

*Psychology only includes 7 topics here as the remainder occurred in two degrees or fewer and/or represented less than 1% of the total topics in the discipline, and therefore were ineligible for *MTW* scores

Through visual analysis of this list, some degree of ‘face validity’ is given to the approach; the topics highlighted in each discipline do display relatively strong association with their fields. However, it is important to note that there is a gap between these and the most frequent topics in each discipline (or indeed the topics the casual observer may most associate with each). One obvious example is OR and “optimisation”. This topic is indeed the most frequent of all in the discipline, and also the one many will most closely associate with OR. However, in figure 2 this is shown to be only 7th in terms of ‘importance’ using this weighting scheme. The cause is that “optimisation” also features at least

once in each other discipline category except for psychology. In other words, if an analytics degree features an “optimisation” module we could not be completely certain that this is indicative of an OR association, merely that there is a strong probability, something which the *MTW* weighting reflects.

Using these code frequencies, we are able to compute a total score for each discipline; that is the sum of the frequency of all topics in the full analytics corpus multiplied by the individual weightings of that term in each discipline (the majority of which were zero for the reasons given above). In other words, we provide a probabilistic judgement (based on *MTW*), as to the extent to which the analytics modules relate to the ‘traditional’ disciplines included in the study. This allows for a comparison of these totals across disciplines, an indicator of the relevance of each in the curricula of analytics degrees, as shown in figure 3.

Figure 3 Summed Module Scores by Discipline (Analytics Degrees)

As with the ensemble model of section 4.1, OR remains prominent. In contrast, however, IS is the second most prominent, with machine learning only fourth, scoring marginally below statistics. Psychology scores very lowly, not completely surprising in terms of *a priori* theory, and somewhat supported by only one degree being classed with this label. However, whilst these results are useful, the indication of the previous analysis is that, to some degree, this may be comparing apples and oranges, in that the main analytics degrees incorporate two separate categories; those classed as machine learning and another as OR. Therefore, we performed the same procedure on each of these subsets (separating based on classification) to analyse the module weightings associated with each.

Despite the lower *overall* score of machine learning, obviously it would be expected that the discipline would score better in the subset of degrees classified as “ML” in the ensemble model. As shown in figure 4, the results on this subset alone, this indeed was the case. Machine learning is now by far the most prominent, with computer science the second highest scored. OR is far less

influential in this dataset as the second smallest, with statistics taking a significantly lower position than with the totals. IS remains reasonably prominent as the third highest in terms of module topic scores.

Figure 4 Summed Module Scores by Discipline (ML Classified Degrees)

Finally we analysed the subset that were classified as “OR”, shown in figure 5. As would be expected, OR is the most prominent discipline in this dataset, followed by IS and then statistics. Machine learning and computer science drop the most in comparison to figure 4, now ranking 4th and 5th respectively.

Figure 5 Summed Module Scores by Discipline (OR Classified Degrees)

To further analyse each, the top 20 topics in each subset are reviewed, as shown in figure 6. The first category displays, unsurprisingly, a close association to machine learning, as well as related fields such as big data, web and eBusiness, and data mining and specific techniques such as visualisation and distributed/high-performance computing. Additionally, there is a clear emphasis on other

computer science-type topics including software, human-computer interaction (HCI), databases and cloud computing. The second category features many of the topics prominent in OR courses such as forecasting, optimisation and simulation, whilst statistics and OR as a whole are listed as the second most frequent topics. Additionally, there is a very clear business theme with a variety of domains including marketing (the top by some margin) as well as more general business topics such as logistics and supply chain management (SCM), management (as a whole), finance, strategy and operations management. As discussed, *decision making* modules are more prevalent here, with decision sciences and consulting key topics (additional to the domain specific terms already discussed).

Figure 6 Top 20 Topics in the Two Categories of Analytics Degrees (by Frequency)

First Category ("ML" classed degrees)			Second Category ("OR" classed degrees)		
Rank	Module Topic	Freq.	Rank	Module Topic	Freq.
1	Big data	12	1	Marketing	21
2	Programming	11.5	2	Operational research	8.5
3	Data analysis	9.5	=	Statistics	8.5
4	Web & eBusiness	9	4	Logistics & SCM	8
5	Data mining	8	5	Management	7
=	Machine learning	8	6	Data mining	6.5
7	Visualisation	7	=	Decision sciences	6.5
8	Distributed computing	6.5	8	Forecasting	5.5
9	Statistics	6	9	Consulting	5
10	High performance computing	5	=	Data analysis	5
=	Optimisation	5	11	Finance	4.5
=	Software	5	=	Programming	4.5
13	Cloud computing	4.5	13	Big data	4
=	Data management	4.5	=	Natural language processing	4
=	Databases	4.5	15	Economics & econometrics	3.5
=	Operational research	4.5	=	Operations management	3.5
17	Economics & econometrics	3.5	17	Algorithms & complexity	3
=	HCI & usability	3.5	=	ERP	3
=	Information retrieval	3.5	=	Information systems	3
=	Natural language processing	3.5	=	Web & eBusiness	3
			=	Optimisation	3
			=	Project management	3
			=	Revenue management	3
			=	Simulation	3
			=	Spreadsheets	3
			=	Strategy	3

5.1 Summary

The module analysis presented in this section was designed to meet RO2: *to determine the specific skills, subjects and techniques taught within analytics degrees*. However, following on from the findings of the previous section (classification of degrees), this analysis is seemingly more meaningful when applied to the two main categories of analytics degree found in the earlier analysis; the

degrees which most closely align to machine learning, and to OR respectively. The two categories demonstrate a relatively ‘clean’ separation. The skills and techniques most frequently taught in each not only align to these disciplines (machine learning and OR), but also demonstrate other unique characteristics such as an association with computing in the case of the former, and business topics with the latter. Such a separation provides further validation to the ensemble classifier model, but also provides further detail on the different characteristics of each, and a more granular perspective on the specific approaches common to each. This line of thought is concluded in the following section, where a more complete description of the two ‘types’ of analytics degree these analyses suggest.

6. A Typology of Analytics Education

The analyses, presented in sections four and five, suggests that, in combination with some subset areas such as courses that are more management orientated (psychology) or technology management (information systems), and indeed other specialisations would be feasible, analytics Masters degrees in the UK broadly fit into two categories. The first category is most closely aligned to machine learning, primarily emerges from computing and technology schools, whereas the second category, aligned to OR, will typically be based in business schools.

These findings are supported by analysis of universities that offer more than one degree in the analytics area. In total 7 universities fall into this category as shown in table 7. In many (four of the seven) of these cases there is a clear ‘two-pronged’ approach, with a business school-based course classified as OR, often named a variant on “business analytics”; and a course with a classification of machine learning, based in computing schools, and typically with “data” in its name.

Table 7 Universities with Multiple Analytics-Type Degrees

University	Degree Title	School	Class
University of Edinburgh	Marketing & Business Analytics	Business School	OR
University of Edinburgh	High Performance Computing with Data Science	Parallel Computing Centre	ML
University of Essex	Business Analytics	Business School	OR
University of Essex	Big Data and Text Analytics	Computer Science & Electrical Engineering	CS
University of Essex	Data Science	Computer Science & Electrical Engineering	OR
Lancaster University	Data Science	Science & Technology	OR
Lancaster University	Management Science and Marketing Analytics	Management School	OR
University of Leeds	Business Analytics and Decision Sciences	Business School	OR
University of Leeds	Advanced Computer Science (Data Analytics)	Computing	ML
Swansea University	Finance and Business Analytics	School of Management	ST
Swansea University	Management (Business Analytics)	School of Management	PS
UCL	Business Analytics	Computer Science & Management Science	OR
UCL	Web Science & Big Data Analytics	Computer Science	ML
University of Warwick	Business Analytics	Business School	OR
University of Warwick	Data Analytics	Computer Science	ML

There are several possible reasons as to why this separation may occur. In an ideal world the assumption would be that the different degree categories have emerged to meet specific training needs, such as those discussed at the start of the paper. However, it also appropriate to acknowledge that there may be more pragmatic reasons behind this. The provision of Masters degrees is obviously dependent on teaching resources, and it is obviously easy, quicker and less costly for a university to utilise existing staff to this end. In the case of the creation of analytics degrees, if there are staff already employed with experience in areas such as machine learning and OR, this could explain why degrees come to take these attributes. Secondly, most universities will be divided into specific schools and faculties, each of which have their own specialisations and topic boundaries. If, as this research would suggest, analytics has aspects of both computational elements as well as business elements, the school in which the degree is hosted is likely to have strong bearing on which of these orientations is stronger in its curricula.

These two categories of degrees do, however, demonstrate sufficient difference to make their distinction significant, but also suggest some important implications for a wide range of people involved in analytics and related disciplines. The characteristics of each of these are presented as a summary in table 8, whilst the remainder of this paper will consider these implications, and also the contributions of this research.

Table 8 Summary Characteristics of the Two Types of Analytics Degrees

	Type One	Type Two
Likely title(s)	"Data Science" / "Big Data"	"Business Analytics"
Classification	Machine learning	Operational research
Likely school	Computing / technology	Business school
Linked disciplines	Machine learning (29.63%) Computer science (22.91%) Information systems (17.72%)	Operational research (44.88%) Information systems (20.38%) Statistics (18.33%)
Most likely module topics	Big data (6.96%) Programming (6.67%) Data analysis (5.51%) Web & eBusiness (5.22%) Data mining (4.64%) Machine learning (4.64%)	Marketing (13.25%) Operational research (5.36%) Statistics (5.36%) Logistics & SCM (5.05%) Management (4.42%)

7. Contributions and Implications

As alluded to in the introduction, the results of these analyses, suggesting two seemingly quite contrasting types of analytics degrees are currently offered at UK universities, has implications for many involved in analytics. This section will further assess these (thus meeting RO3: to evaluate the current academic response to analytics; the implications this has for employers and educators, and also for the operational research discipline), along with the contributions of the research as a whole.

7.1 Theoretical Contributions

The first theoretical contribution of the research is towards furthering our understanding of the field of analytics, in particular in reference to related disciplines. Our results demonstrate that analytics is inherently multi-disciplinary, as has been suggested elsewhere in the literature, with both analyses indicating that a range of skills and disciplines are involved in their curricula. However, whilst there are some combination of technology, quantitative and decision making disciplines (Mortenson *et al*, 2015) in each, the analyses presented suggests that the combination varies dependent on the 'flavour' of analytics discussed.

Additionally, though more indirectly, the analysis gives some perspective on the nature of academic responses to changing demands in the broader environment. Although no definitive evidence as to why the courses have developed as they have (seemingly into two main 'types' of analytics degree) has been given, there is some inference that there represents something of a trade-off for universities. On the one-hand there is a need to create new materials tailored to the new

requirements of industry; on the over there is the need to utilise current resources (e.g. teaching specialisms), as well as building within current provisions (e.g. existing modules).

Finally, we argue the paper offers contributions in respect to the methods employed. As stated, we believe this is the first research paper in this area to utilise a fully algorithmic approach to such a task (with the ensemble analysis). It is our belief that the utilisation of machine learning methods to analyse online course materials offers many potential opportunities. Such approaches are made more accessible by the big data revolution, obviously an important factor in the growth of analytics, and offers the potential to open new avenues of research and new methods of inquiry for business and social science research. Additionally we have presented a new approach (module topic weighting) that can be used in 'hybrid' quantitative/qualitative research studies, an approach that can control for the duplication of terms across multiple topic categories.

7.2 Practical Implications

Our final contribution we believe to be a practical one; by providing new insights into the variety within analytics training that may be of use to a wide range of people involved in the analytics area. Most obviously, for potential students of such degrees there are implications as to the type of course that would best suit their requirements. This decision may be influenced by building upon their prior experience (if they have a business or computing background); a desire to move from one of these areas to the other; or the branch of analytics they are most interested in moving into.

For employers of these graduates, there is an onus to carefully consider not only the attributes of the candidates they seek, but also the role they perceive analytics to play within the organisation (or at least the analytics function to which they are recruiting). For businesses with considerable data stores from which they wish to extract value, for whom analytics is housed within an IT function, or those which favour decision automation tools and intelligence reporting, type one courses are likely to provide the best candidates. In such cases computing skills will be at a greater premium, as will be the ability to extract information from data (big or otherwise) through data mining and machine learning techniques.

If instead the preference is for an analytics function to work more closely with (or even within) business functions, or businesses who seek analytical solutions for more politically and/or socially complex problems or problems, type two graduates are likely to be a better fit. Modelling problems of this kind are more likely to be collaborative efforts between analyst and subject matter experts, necessitating the candidate to be better versed in business practice and culture. Whilst such analysis will invariably utilise some form of data, in many ways this will be a secondary concern in this form of analytics, with developing an understanding of the specifics of the business problem and

influencing factors likely to be the first, and critical, stage. Thereafter, the use of OR or statistical models are more likely to be applied, as such, approaches more readily allow the modelling of external constraints and concerns, particularly if these are less tangible and subjective.

Finally, in many ways the suitability of the candidate is likely to be a consideration based not only on the technical, but also on the cultural. For a type one analyst being able to operate within an IT-type function will be critical, and this implies an ability to have an understanding of the wider infrastructure and the systems of the organisation, within which their solutions will be housed. For a type two analyst the ability to work collaboratively with a wider range of departments and people, and to develop a broader understanding of the organisations' strategic and environmental concerns, will be integral to the success of their models and analyses.

The implications for academics and university teaching staff are multiple. Firstly, for those involved in operating analytics degrees such as these, it is important to consider where this degree is located within this typology. This impacts on a wide range of concerns ranging from competitor research and benchmarking, to student recruitment and managing student expectations. Secondly the analysis may help course designers to find opportunities for differentiation, and for seeking collaborations within or without their university.

More broadly, and central to one of the core themes of this research, the results provide new insight into the academic traditions from which analytics draws. The prominence of machine learning is perhaps unsurprising considering the focus and coverage the discipline has recently been afforded. The impact of statistics is, arguably, lower than may have been expected, as the discipline concerned primarily with data and its analysis, but it seemingly has an important role particularly in type two degrees, and many statistical topics were highlighted in the module analysis. Computer science has a similarly strong influence in the modules of type one degrees (those classed as machine learning). Psychology is seemingly the least important, however, its inclusion does help highlight the role of *decision making*, particularly in the degrees that combine analytics with management studies.

For the IS discipline, as the third most frequent classification (albeit some distance behind the top two) it would appear to have an important role to play. IS topics were the third and second (respectively) most important in the weighted scores for *both* of the main types, and was second highest in the overall scoring. This suggests that the IS discipline may offer an important bridge between the two types of analytics degrees presented. In that IS, following on from the taxonomy of figure 1, is conceived to have both *technological* and *decision making* elements, for businesses that engage both type one *and* type two models of analytics, IS graduates may be critical in enabling these two to interact effectively. Alternatively, IS training may be similarly influential in connecting

technology orientated analytics functions (type one) to the business; or business orientated analytics functions (type two) to the IT infrastructure of the organisation.

As this discussion helps highlight, analytics degrees do indeed draw from many disciplines, and therefore would be of concern for many different academic communities. However, considering the results of the first analysis, where nearly half of the degrees were classed as “OR”, the implications may be most significant for the operational research community. In and of itself this finding is a notable one, as it suggests the important role the discipline plays within analytics education. This also adds weight to the suggestion of analytics as a significant opportunity for the OR community (e.g. Liberatore and Luo, 2010). However, there are other elements of these findings that may present concerns for the community.

Firstly is the volume of degrees classified as machine learning at two higher than OR. Further to this, the study presented here essentially seeks to assess *academic provision*; there is no assessment of the relative demand in the jobs market for graduates in these different forms of analytics degree. The evidence presented here, as well as other examples of the use of machine learning techniques in business analytics, should promote some debate as to the extent of the role machine learning techniques may play in OR practice.

Ultimately, as the first discussion of these two classifications of analytics degrees, there remains significant uncertainty as to the role OR may play in the future development of analytics, and the relative demand for staff in type one or type two analytics graduates. It is the argument of this paper that this should represent a significant concern for the OR community, and as such an area we would recommend future research in.

7.3 Summary

Over the course of this section we have demonstrated the theoretical contributions this work has provided, as well as the implications our findings have for the analytics community, and, most significantly, the OR community. In doing so we have met the final research objective (RO3). The paper will conclude with a brief summary of the work and suggested limitations and future research.

8. Concluding remarks

This paper has presented an examination of the contents of UK Masters courses through two analyses. We have identified two, seemingly distinct and diverse approaches to analytics education, thus meeting RO1. Secondly we have analysed the most frequent modules across all analytics degrees, as well as those most frequently occurring in each of the approaches above. This analysis has further demonstrated differences between the categories discussed (RO2). Finally we have

discussed the implications this has on employers and educators, as well as the OR community at large (RO3).

There are, of course, limitations and possible directions for further research. Firstly, this study has focused solely on degrees from the UK. Whilst, with the increasing globalisation of the higher education industry (Masters degrees in particular), it may not be unreasonable to consider this as a suitable indicator for global trends, at the same time further research could use similar methods to examine degree materials from the US and other parts of the world. Secondly, there are some question marks around the statistical validity of some of the measures used, with a relatively low accuracy score for the ensemble model and the lack of hypothesis testing in the module analysis. This (arguably) is attributable to the variable data quality for the former, and the ratio of data points to variables in the latter, and we would argue that through the use of multiple methods we have provided extra validity to these results. However, over a larger scale study, possibly internationally, some of the issues may be countered (as it may remove the need for using job advert data as a proxy for course materials in the ensemble, and potentially allow for significance tests to be applied to the module analysis).

In summary, whilst analytics education remains still at a nascent stage, already clear patterns have emerged. With big data and an ever increasing demand for skilled professionals seemingly here to stay, the development of these courses will be a continued concern. For the OR community in particular, significant issues remain as to whether the type two degrees, to which OR is more closely aligned, will remain as the more prominent of the two, or indeed which are more in demand from analytics employers. More generally, there remains a need for the disciplines aligned with the field, OR included, to adapt to ensure their graduates develop the relevant skills and knowledge to succeed in this changing business environment. Reflection and adaption will be critical to the success of any practically-orientated discipline, whether they be directly producing the analysts working at the technical end of analytics, or graduates who utilise and embed analytics in their organisations.

Acknowledgements and Further Technical Details

The authors acknowledge the significant contribution to this paper from their late colleague and co-author, Professor Neil Doherty. The authors would also like to acknowledge the support of the Operational Research Society who part-funded and supported this research, as part of the charitable project titled *Is Operational Research in UK Universities 'Fit-for-Purpose' for the Growing Field of Analytics?*

For any readers interested in applying the methods used in this paper, a technical addendum has been produced, including a list of all steps and source code where available that we are happy to

share with any interested parties. Please contact the authors using the correspondence email address given at the start of the paper.

References

- Aizerman MA, Braverman EM and Rozonoer LI (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* **25**: 821–837.
- Campello RJGB, Moulavi D and Sander J (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G (eds). *Advances in Knowledge Discovery and Data Mining*. PAKDD 2013. Lecture Notes in Computer Science, vol. 7819. Springer, Berlin, Heidelberg
- Chiang RHL, Goes P and Stohr EA (2012). Business intelligence and analytics education, and program development: A unique opportunity for the Information Systems discipline. *ACM Transactions on Management Information Systems* **3**: 12-24.
- Chu H (2006). Curricula of LIS programs in the USA: A content analysis. In: Khoo C, Singh D and Chaudhry AS (eds.), *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006*, 3-6 April 2006. Singapore: Nanyang Technological University, pp. 328-337.
- Cortes C and Vapnik V (1995). Support-vector networks. *Machine Learning* **20**: 273-297.
- Cox D (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, **20**: 215-242.
- Danson F, Amjad A, Davenport T, Gold S, Guszczka J, Katyal V, Lucker J, Mueller B and Raphael J (2016). *Analytics trends 2016: The next evolution*.
<http://www2.deloitte.com/content/dam/html/us/analytics-trends/2016-analytics-trends/pdf/analytics-trends.pdf> (accessed February 2016).
- e-Skills and SAS (2013). *Big data analytics: Adoption and employment trends, 2012-2017*.
<http://www.sas.com/offices/europe/uk/downloads/bigdata/eskills/eskills.pdf> (accessed July 2014).
- Ester M, Kriegel HP, Sander J and Xu X (1996). A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Knowledge Discovery in Databases*, **96**: 226-231.
- Evans JR (2012). Business analytics: The next frontier for decision sciences. *Decision Line* **43**:
[http://www.decisionsciences.org/DecisionLine/Vol43/43_2/dsi-dl43\(2\)-mar12-final.pdf](http://www.decisionsciences.org/DecisionLine/Vol43/43_2/dsi-dl43(2)-mar12-final.pdf) (accessed June 2012).
- Fan R-E, Chang K-W, Hsieh C-J, Wang X-R and Lin C-J (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* **9**: 1871-1874.
- Fukunaga K and Narendra PM (1975). A Branch and Bound Algorithm for Computing k-Nearest Neighbors. *IEEE Transactions on Computers*, **100**: 750-753.
- He H and Garcia EA (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**: 1263-1284.
- HM Government (2013). *Seizing the data opportunity: A strategy for UK data capability*.
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf (accessed June 2015).
- Ho, TK (1995). Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Quebec, Canada, 1995, pp. 278-282.
- IBM Corporation (2012). *The IBM Big Data Platform*.
<http://www.maximesolutions.co.uk/Resources/The%20Big%20Data%20Platform.pdf> (accessed December 2015).
- LeCun Y, Bottou L, Orr G and Muller M (1998). Efficient backprop. In Orr G and Muller K (eds.) *Neural networks: Tricks of the trade*. Springer: Berlin (pp. 9-48).
- Liberatore M and Luo W (2010). The analytics movement: Implications for operations research. *Interfaces* **40**: 313-324.
- Maaten LVD and Hinton G (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**: 2579-2605.
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C and Hung Byers A (2011). *Big data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute.
http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation (accessed April 2014).

- Marr B (2018). *How Much Data Do We Create Every Day?* Forbes.
<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/> (accessed August 2018).
- McInnes L, Healy J and Astel S (2016). HDBSCAN. <https://hdbscan.readthedocs.io/> (accessed May 2018).
- Molluzzo JC, Lawler JP and Plaza OP (2015). A proposed concentration curriculum design for big data analytics for information systems students. *Information Systems Education Journal* **13**: 45-57.
- Mortenson MJ, Doherty NF and Robinson S (2015). Operational research from Taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research* **241**: 583-595.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**: 2825 - 2830.
- PyData Development Team (2012). Pandas. <http://pandas.pydata.org/index.html> (accessed March 2015).
- Rae S (2018). *Big Data Skills Shortages – and how to work around them*. Computer Weekly.
<https://www.computerweekly.com/opinion/Big-data-skills-shortages-and-how-to-work-around-them> (accessed August 2018).
- Renders J-M (2004). Kernel methods for natural language processing. In: *Learning methods for text understanding and mining*. PASCAL. January 26-29, 2004. Grenoble: Xerox Research Centre Europe.
- Salton G (1991). Developments in automatic text retrieval. *Science* 30 August 1991: 974-980.
- Stern DT (1998). Practicing what we preach? An analysis of the curriculum of values in medical education. *The American Journal of Medicine* **104**: 569-575.
- van der Walt S, Colbert SC and Varoquaux G (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering* **13**: 22-30.
- Varshney KR and Mojsilovic A (2011). Business analytics based on financial time series. *IEEE Signal Processing Magazine* **28**: 83-93.
- Ward JH (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association*, **58**: 236-244.
- Wixom B, Ariyachandra T, Goul M, Gray P, Kulkarni U and Phillips-Wren G (2011). The current state of business intelligence in academia. *Communications of the Association for Information System* **29**: 299-312.
- Wolpert DH and Macready WG (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, **1**: 67-82.
- Zhou Z-H (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press: Boca Raton, FL.

Appendix

Data cleaning and transformations

By default, text data, as an unstructured dataset, requires several pre-processing steps and decisions need to be made in order to best utilise the resource. The initial task was to control for duplicate data in the job adverts, a common issue in all job posting sites. Firstly all obvious duplications (i.e. those sharing the same job ID) were removed. Secondly we randomly selected the job adverts to use in the models, excluding up to 50% dependent on the quantities extracted. Finally, we sought to identify and remove adverts that were *likely* to be duplicates. To do so we used the Python package *difflib* to identify similarities in word usage between two text objects, returned as a decimal between 0 and 1, where 1 represents an exact replicate. Accordingly we excluded any document with a similarity rating (α) above 0.7 with any other advert, an arbitrary figure essentially, but one which performed well during experimentation.

Secondly, we removed all numbers, HTML code, punctuation, and any word of fewer than 3 characters. Thirdly we applied stemming. Common to such applications, stemming involves reducing words to their shortest 'stem' so that different variations of the same word are treated the same (e.g. different tenses and pluralisations of the same words are converted to stems such that the model sees them as the same). As example, the words "businesses" and "businessman" would both be changed to the stem "busi". Having applied this transformation we then sought to eliminate the biases in the data introduced by highly frequent terms. Words such as "I", "you" or "and" are common to all documents, yet offer little in terms of meaning were removed. Similarly, other very frequently occurring words across a corpora were removed, as again they offer little actual value in understanding the documents. Finally the words that occur very infrequently were removed as they have little predictive value, but conversely add significant computational complexity.

In a similar vein we sought to reduce complexity by transforming the documents into a document-term-matrix (DTM); that is a $N \times M$ matrix where the rows are the documents $1:N$, the columns are the words in the vocabulary $1:M$, and the contents are the relative frequencies of these words. Such a transformation means that the order of the words in the document are lost, and instead they are treated as a 'bag-of-words'. This obviously results in the loss of some meaning as a trade-off for reduced computational complexity. On the surface this may be of particular relevance, as LinkedIn do use headers in their job adverts (e.g. "skills and experience"), and course materials may include headings such as "module", both of which may give extra weight to words included in these sections. However, the usage of these headings varies greatly such that the value in retaining them is fairly limited.

Lastly, for similar reasons to those given in section 3.2, we sought to transform the data from simple counts into scores that reflect the relative importance of the term to each document. This was performed using the term frequency-inverse document frequency (TF-IDF) algorithm (Salton, 1991). The algorithm effectively applies a penalty to terms that are frequent across many documents in the collection such that the terms that score highest will be those that are relatively frequent in a document, but not in the majority of documents. In other words, terms such as "degree" or "position", which will be frequent in most university course materials and job adverts respectively, will be penalised as they reveal little about the differences between different sources. Contrastingly, words which are infrequent in other documents, but relatively frequent in some, will be promoted. For example, a term such as "Bayesian", which will be typical of statistics-orientated degrees or jobs but infrequent in others, will be scored highly.