Economic and Social Research Council
North West Social Science Doctoral Training Partnership

NWSSDTP Specialised Training Routes

The ESRC have awarded the NWSSDTP 12 steered awards each year in the following Priority Areas:

- Advanced Quantitative Methods (4 awards)
- Use of ESRC data sets (4 awards)
- Interdisciplinary research, which straddles other research council remits (4 awards)

To encourage full take up of these awards, the NWSSDTP is able to offer enhanced stipends of £3000 per annum to students whose projects align with the following Specialised Training Routes identified by the ESRC as strengths within the consortium, which fit within the Priority Areas above:

- Advanced Quantitative Methods (AQM)
- Longitudinal Data Analysis (for Use of ESRC Data sets)
- Data Science (for Interdisciplinary research, which straddles other research council remits)

To apply for a Specialised Training Route enhanced stipend, students need to complete the relevant part of the Priority Areas and Specialised Training Routes section of the Studentship Application Form. The award of these enhanced stipends will be competitive. The relevant section of the guidance below should be read carefully prior to completing this question.

We also encourage students to identify if their proposed project fits into a Priority Area, but does not fall in one of the Specialised Training Routes outlined above. In these cases, if the candidate is successful an enhanced stipend will not be payable, but the application will be considered positively in terms of fit with a priority area.

What is AQM?

The expectation is that an AQM student/proposal would meet all of the criteria below:

1. Methodological contribution criterion
Will the proposed research produce a contribution to quantitative methodology i.e. does it go beyond simply applying standard quantitative methods to a particular substantive research problem? This is not to say that only methodological proposals will be considered but that at least part of contribution to knowledge that the proposed research would provide would be methodological. So, for example, the application of cutting edge statistical or mathematical analyses that examine sensitivity to assumptions about missing data, measurement error, etc. would be within the scope of this criterion as would a methodological analysis of the application of an advanced quantitative method to a particular data configuration or an applied piece of research with any of the above methodological contributions as a means to an end. When considering the relative merits of proposals (of equal quality) that meet these criteria, those that propose genuine methodological innovation would be favoured over those that simply apply advanced methods.

2. Supervisor Expertise criterion
At least one member of the proposed supervision team should have specialist expertise in advanced quantitative methods (defined by them meeting criteria 1 and 3 in their own research).

3. Publication criterion

Is the proposed research of type which could **in principle** be published in a journal which **specialises** in publishing articles using or researching advanced quantitative methods? Examples of such journals are:

- Journal of the Royal Statistical Society Series A B or C
- Journal of the American Statistical Association
- Annals of Statistics
- Survey Methodology
- Demography
- Journal of Econometrics
- Econometric Theory
- Journal of Applied Econometrics
- Econometrica
- Quantitative Econometrics
- Biometrika
- British Journal of Mathematical and Statistical Psychology
- International Journal of Geographical Information Science
- Geographical Analysis
- Journal of Regional Science
- Journal of Empirical Finance
- Journal of Financial Econometrics
- The Journal of Quantitative Criminology

Note that this list is not meant to be exhaustive but provides an indicative cross-disciplinary range.

Are the criteria relative or absolute?

One key question is whether the criteria should be applied in an absolute sense or relative to disciplinary norms. There are advantages and disadvantages to both approaches. A fully relative system would have a tendency to disincentivise applications to disciplines where advanced quantitative methods are central. On the other hand a fully absolute system may exclude some worthwhile applications in disciplines which are not normally quantitative. Both these extremes would subvert the strategic aims of the AQM scheme.

The approach we have adopted is to work with a set of absolute standards but then to allow some latitude for applications from non-quantitative disciplines. However, irrespective of disciplinary context, proposals that simply apply quantitative methods set out by ESRC as fundamental expectations of basic social science training will not, be regarded as AQM.

Does AQM mean high quality?

It would be surprising if a poor quality application was classified as AQM. However, the judgement of the quality of an application is largely independent of whether it meets the AQM criteria above.

Entailments of being an AQM student

1. AQM students will be required to submit an annual report form at the end of years 1 and 2 to the AQM sub-committee. The report will comprise three sections: research undertaken (and planned), training undertaken (and planned) and outputs (and planned).This sub-committee will review the annual report to confirm that the student/research still meets the criteria for AQM and will provide recommendations to the full DTC committee regarding continuation or not of the stipend.

2. AQM students are expected to be a member of the AQM student community. It is expected that this will minimally entail attendance at 3 one day workshops (one at each institution) per year and to make presentations at least 2 during the lifetime of their studentship.

What is Longitudinal Data Analysis?

Introduction

This specialised training route forms part of the Priority Area focusing on the use of ESRC datasets. Despite the name of the Priority Area, applications do not necessarily need to use ESRC datasets, and other longitudinal datasets will be accepted. It will support students to develop appropriate methodological skills to study change over time within individuals or subjects

The characteristics of longitudinal data which distinguishes it from time series data is that in general there are a large number of distinct subjects (N large), covariates (time stable and possibly time varying) are collected and assumed to be associated with change over time, and the number of time points on each individual is small compared to the number of individuals.

The expectation is that students on this training route will be using secondary data sources. While time is a key component of social science research and many research studies involve the collection of data at different time points or observing and particular behaviours change over a relatively short period of time, the emphasis of this training route is towards longer time periods,

There are two main types of study which will be considered under this training route.

Analysis of longitudinal Social Survey data

- Studies will involve secondary analysis of existing data. While most applications will be using ESRC longitudinal datasets, there are possibility for using longitudinal survey data from other countries. A list of ESRC longitudinal data sets is in Appendix 1.
- Data will be taken from at least two time points with a minimum of one year between these time points.
- Individuals sampled a time 1 should be followed up at all subsequent time points, subject to attrition and refreshing of the sample
- There is no maximum number of time observations
- Data may be either qualitative of quantitative

Analysis of event data and longitudinal data arising from administrative data sources

Longitudinal data may also consist of data on events, and will arise naturally from administrative data. For example, data on hospital patient episodes, residential mobility, job mobility or police arrests over a sample of individuals will give a variable number of observations per subject, and the time between successive events will be unique to each individual. The time of events will be recorded (date-stamped) along with other information on the events, and time constant and possibly time varying covariates.

What is not in this specialised training route

Projects funded under this specialised training route will not

- Carry out primary data collection
- Involve experimental rather than observational designs
- examine very short changes over time (such as daily or hourly stock market movements)

Students are not restricted to analysis of quantitative data under this specialised training route and where appropriate are encouraged to consider the possibility of using longitudinal *qualitative* data. This data will again be defined as data collected from the same individuals with at least two time points with a minimum of a year in between data collection. The ESRC Timescapes project supports longitudinal qualitative data and cohort studies, including NCDS and ELSA, also include qualitative data.

Candidates applying for the specialised training route in longitudinal analysis should state in their application the datasets that they intend to use and provide a brief justification of how their data analysis meets the criteria of longitudinal analysis. For example 'using Waves 1 and 4 of Understanding Society to explore changes in well-being over time'.

ESRC-funded longitudinal datasets
The ESRC-funded longitudinal datasets are:

- British Household Panel Survey
- Understanding Society
- 1958 National Child Development Survey
- British Cohort Study 1970
- Next Steps (Longitudinal Study of Young Persons in England)
- Millennium Cohort Study
- UK Census Longitudinal Studies
- English Longitudinal Study of Ageing
- Avon Longitudinal Study of Parents and Children
- Timescapes (qualitative data)

What is Data Science?

A social science perspective on data science
Data science is best understood as a methodological framework for collecting, manipulating, analysing, transforming, visualising and representing data. The framework is interdisciplinary; drawing heavily - but not exclusively - on computer science and statistics.

The advent of so called big data signalled a blossoming of data science. Dealing with volume, velocity and variety (the supposed defining features of big data) are key data problems where data science comes into its own. However, data science is not only about big data. The term can be applied to any data problem where an adaptive analytical approach is either necessary or simply beneficial, although conventionally data science research involves challenging computing to collect, store or manipulate the data, or novel inferential approaches to accommodate challenging data questions.

Some critics of data science respond to the adaptive analytical approach by criticising it as atheoretical. However, whilst it is true that some data science is purely inductive (the often cited Google flu trends being a good example), this is not a defining feature; nor one which has general support within the social sciences (Kitchin, 2015). However, it is perfectly plausible, and we believe desirable, to utilise data science in a theory led manner. Indeed, in common with much well conducted social science, data science practice is often a fusion of the deductive and the inductive, with the information and knowledge that has been extracted from data being used to develop theory and theory in turn shaping our explorations of data.

One thing that is often not clear to an outsider is how data science is different from statistics, or indeed quantitative methods. This is where the computer science – or perhaps more precisely informatics -

element is vital. To paraphrase Diggle (2015); the statistics component of data science enables us to understand data whereas the computer science component enables us to make data understandable. Thus, in addition to the inference component of an analysis, data science provides methods to gather, store and pre-process data prior to (and during) inference, and to facilitate action on the results of the inference. This also hints at a further facet of data science which is the strong overlap to the expansive toolkit of computer vision research, which has developed visual analysis and representation techniques that are more appropriate to those highly dimensional, less structures and more uncertain data that are typically commonly used in the practice of data science.

Our view is that data science and the social sciences are in fact actually naturally aligned, and indeed complementary. We like to think that, rather than being theory or data led, data science is best thought of as curiosity led; one image is of a data scientist shaking an enormous tin of apparently murky data sludge and seeing what drops out. Social science data, whether it is at the detailed micro level of experimental psychology or at the large scale level of socioeconomic or social media data streams, is inherently murky, dirty, ambiguous, contradictory and confounding. As a consequence social science (and related fields) is an imperfect science – it does not quite meet the stringent Popperian criteria of falsifiability; Temple (2012). However, this is precisely the sort of data and the sort of epistemological fuzziness that data science is made for. In short data science has the potential to make the social sciences fulfil their huge potential.

Beyond this methodological mapping, there are further fruitful areas where social science has a significant role to play in the wider development of data science. For example, in the ethics of new forms of data and analysis, recognising that most if not all data cannot be isolated from those social processes that govern their collection; or, in those social considerations that surround the proliferation of algorithmic governance associated with the management and planning of many established and new urban areas (e.g. Future Cities). We would consider such projects would be within in scope and would seek to include a "social science of data" strand in the PhD projects funded under the proposed scheme.

What should a data science proposal look like?
The expectation is that a Data Science student/proposal would meet all of the criteria and would fulfil both of the entailments.

Methodological contribution criterion
Will the proposed research produce a contribution to data science methodology i.e. does it go beyond simply applying routine methods to a particular substantive research problem? This is not to say that only methodological proposals will be considered but that at least part of contribution to knowledge that the proposed research would provide would be methodological. So, for example, the application of cutting edge machine learning techniques would be within the scope of this criterion as would a methodological analysis of the application of data science methodology to a particular data configuration or an applied piece of research with any of the above methodological contributions as a means to an end. When considering the relative merits of proposals (of equal quality) that meet these criteria, those that propose genuine methodological innovation would be favoured over those that simply use data science methods.

Publication criterion
Is the proposed research of type which could in principle be published in a journal which specialises in or frequently publishes articles using or researching data science? Examples of such journals are:

- EPJ Data Science
- Information Fusion
- Big Data

- Big Data Research
- Intelligent Data Analysis
- International Journal of Data Science and Analytics
- Journal of Big Data
- Machine Learning
- Journal of Mathematical Sociology
- Social Science Computer Review
- Sociological Methods & Research
- Journal of the Royal Statistical Society
- Computational Linguistics
- Journal of the American Statistical Association

Note that this list is not meant to be exhaustive but provides an indicative range. It is acknowledged that given data science's interdisciplinary nature, individual students may in fact publish within journals outside of a data science list. The question is not where the students will publish but whether the research could be published in data science oriented journals.

Supervisor expertise criterion
At least one member of the proposed supervision team should have specialist expertise in Data Science (defined by them meeting criteria 1 and 2 in their own research).

Are data science students expected to be based in particular disciplines?
No. Data science is emerging a fully cross-disciplinary methodological framework and therefore data science students could be located within any discipline.

Entailments of being an data science student

1. Data science students will be required to submit an annual report form at the end of years 1 and 2 to the data science sub-committee. The report will comprise three sections: research undertaken (and planned), training undertaken (and planned) and outputs (and planned).This sub-committee will review the annual report to confirm that the student/research still meets the criteria for data science and will provide recommendations to the full NWSSDTP committee regarding continuation or not of the stipend.
2. Data science students are expected to be a member of the Northern data science student community. This community will also involve students within the Data Science and Society CDT and some from the white rose DTP. It is expected that this will minimally entail participation at one event each year.

How to apply for a data science stipend
If you think your application is eligible for data science simply check the data science box on your NWSSDTP application form; the application will then be automatically considered by the data science sub-committee.

The procedure for selection of data science studentships
Prior to the meeting of the main DTC selection committee, the data science sub-committee will meet. The sub-committee consists of 4-5 members of staff from across the NWSSDTP with expertise in data Science and experience as PhD supervisors. The committee will consider all applications for the data science stipend and will (i) decide which applications are eligible for data science and (ii) generate a provisional ranking of those which are eligible on the basis of their methodological strength. The provisional ranking will be advisory only and will be used by the main NWSSDTP selection committee to inform their overall decision making about the allocation of awards.