Data Scientist Internship Programme

Call for proposals and applicant guidance

Application deadline: 5pm Thursday 30th April 2020

Summary

The Leeds Institute for Data Analytics (LIDA) brings together applied research groups and data scientists from a range of disciplinary backgrounds, all of whom are working with data and the latest technologies to generate exciting new evidence on human health, business, science and society. The Data Scientist Internship Programme aims to develop data science excellence within a dynamic and engaging environment; to foster innovation and joint working across University faculties and departments; and to provide interns with the opportunity to apply complex analytical techniques to real data in a bid to solve real-world problems. Now in its 5th year, this Programme has seen successes in the form of an international award from the IALEIA given to one intern for her outstanding analysis of online hate crime data, as well as industry partnerships ranging from Improbable, driving advances in virtual-world development; to global fashion brand, Burberry.

Proposals for intern projects are welcomed from all academic disciplines and partners, and inter-disciplinarity is particularly encouraged: for example, a project may be co-supervised by academics in different Schools or an academic and an external partner. Please note that, so as not to incur delays to the Programme commencing on time, the proposed work must be based on datasets that are already (or soon to be) accessible to the applicant. Applicants are invited to submit proposals of 24 weeks in duration with a start date of either Monday 5th October 2020 or Monday 5th April 2021.

While each project will vary, interns will broadly be expected to:

- Undertake quantitative analysis on core datasets;
- Plan and manage their own research activity in collaboration with the project investigators and other members of the research team;
- Meet with clients and partners in support of LIDA’s partnership-building efforts;
- Participate in LIDA research meetings, events and seminars;
- Generate outputs, including research papers, case studies and other materials for publication.

Interns will be expected to spend the majority of their time at LIDA; however, time working at other Schools or Institutes and/or the premises of project partners is permissible to broaden the intern’s exposure to other networks and ways of working. Interns will typically spend an average of 32 hours per week working on project-related activities, as agreed with their academic supervisors, with a remaining 3 hours per week allocated to training and professional development activities.
A summary of previous intern projects is provided in Annex 1 and short written summaries of a selection of projects are available here.

**Intern Specification**

All interns will be employed by the University of Leeds for a period of twelve months; recruitment for cohort 5 of the Programme will begin in June 2020 so that interns are ready to start at LIDA in October 2020.

The following essential and desirable criteria will be used during the recruitment of the LIDA Data Scientist interns:

**Essential**
- Interns will have excellent all-round numerical skills that may have been gained in the context of quantitative social science, computing, mathematics/statistics, medical bioinformatics or population studies;
- Interns will have completed an Undergraduate or Masters degree in a relevant field;
- Interns will be passionate about using data to answer real-life questions using statistical or predictive techniques;
- Interns will have good communication skills (verbal and written) and a willingness to work as part of a team.

**Desirable**
- Interns may have experience of working with large unstructured data sets;
- Interns may have knowledge of data science, machine learning or data mining;
- Interns may have knowledge of a programming language and/or software for statistical computing;
- Interns may have a PhD in the field of quantitative social science, computing, mathematics/statistics, medical bioinformatics or population studies.

Interns will be allocated to individual projects by LIDA with input from project supervisors.

**Supervision**

Applicants will be expected to supervise the work of the intern throughout the project period; this includes meeting regularly with your intern to manage progress (at least 1 hr per week contact time); brokering meetings between your intern and external partner; and attending review meetings with your intern and LIDA representatives. It is also expected that you will commit as far as practicable to attending presentations delivered by your intern and the rest of the cohort.

**Project duration and start dates**

Projects will last for 24 weeks with a start date of Monday 5th October 2020 or Monday 5th April 2021. Interns will be expected to allocate an average of 32 hours per week to project-related activities, as agreed with their academic supervisors, with a remaining 3 hours per week allocated to training and professional development activities.

**Reporting requirements**

Interns will be expected to report on progress made on a 6-weekly basis via meetings that are either project-specific or group meetings which include all interns. Project-specific meetings will include the following attendees (either in person or via tele- or video conference): the intern; their supervisor; and a representative of LIDA.
Upon project completion, academic supervisors will be expected to provide a written report to LIDA detailing the main outcomes and outputs of the project prior to its end date of 2nd April 2020 or 30th September 2021.

**Eligible costs**

The cost of all intern projects will be £18,500, and this includes:

i) Employment of the intern for a period of 6 months (grade 5, sp 17);

ii) A £1500 training, development and travel budget ring-fenced for use by the intern for their personal development as well as the benefit of the project, in consultation with their project supervisor(s) and LIDA;

iii) All management and administration costs.

This costs breakdown does not include costs relating to support that may be required from the LIDA Data Analytics Team. If you anticipate needing support with data storage, VRE creation, or access to the LIDA safe rooms for your project, you may discuss this and eligible costs by contacting ircdst@leeds.ac.uk, including the following in the subject line – ‘LIDA Intern Project enquiry’.

**Funding & Assessment criteria**

Funding for this call will be provided via one of two routes as outlined below. All applications will be assessed via shortlisting and review by the Project Selection Panel, which will focus on the benefit of the work in general, its technical feasibility and the benefit to the intern (see criteria below).

1. **Self-funded projects**

Project teams may submit proposals that will utilise their own funds, for instance via other university funds, an externally-funded grant, or investment from an external partner. The following funders may be approached depending on relevance of your project to their strategic priorities: EPSRC Impact Acceleration Account (IAA), ESRC Funding Opportunities, Wellcome Trust funding, Science and Technology Facilities Council, and external commercial partners.

Please note – funding allocated from self-funded grants/ budgets must not come from grant monies earmarked for research staff as the interns will be employed under technical contracts.

Assessment criteria:

- Immediate objectives and outcome(s) – has the applicant clearly stated what objectives and outcomes they expect from the proposed project and how the intern will benefit from these?

- Long-term potential – is it clear how the proposed work could lead to further funding opportunities?

- Value for money – is the proposed work proportionate with a 6-month intern project?

- Would the project strengthen LIDA’s research portfolio beyond ESRC & MRC to include the full range of Research Councils and other funders?

- Does the proposal support the Urban Analytics theme in partnership with The Alan Turing Institute?

- Will the project promote interdisciplinary research activities across campus, leading to a grant application and possibly 4*/3* research outputs?

- Will the project develop new data analytics methods and tools for deployment amongst researchers and external partners?

Technical feasibility will be assessed in consultation with members of the Data Analytics Team.
2. LSSI-funded projects

A small proportion of the projects this year will be funded by the Leeds Social Sciences Institute (LSSI) via the ESRC IAA, but please note this funding is finite and will be highly competitive. Project teams should clearly indicate if they are seeking funding via the LSSI in their application, and in these cases applicants must demonstrate how the work is aligned with one or more of the following ESRC IAA strategic priorities:

- Project proposals must align with the ESRC’s disciplinary remit (https://esrc.ukri.org/funding/guidance-for-applicants/is-my-research-suitable-for-esrc-funding/discipline-classifications/);
- The applicant must clearly state what outcomes and impacts they expect from the proposed project;
- The proposal should include engagement with an external partner(s): public, private or third sector;
- It should be clear how, if successful, the proposed work would generate further opportunities to secure funding in the future;
- The proposed work should be proportionate with a 6 month intern project.

Assessment of the merits of the project will be completed by the Project Selection Panel based on the criteria above via a shortlisting process and then decision at the panel meeting.

Technical feasibility will be assessed in consultation with members of the Data Analytics Team.

Data sourcing and handling

The Consumer Data Research Centre (CDRC) is based in LIDA and includes a data hosting service, through which researchers can apply to access a range of open and safeguarded data assets. Details about this service are available at https://data.cdrc.ac.uk. Applicants who wish to discuss this in further detail should contact Paul Evans at p.d.evans@leeds.ac.uk.

In all cases, project data must be confirmed by July 2020. If there is any doubt as to whether data is available beyond this date, LIDA reserves the right to allocate an intern to a different project.

Project proposal submission

Applicants should complete the project proposal form within Annex 2.

Applications should be submitted by email to Kylie Norman at k.r.norman@leeds.ac.uk by 5pm on Thursday 30th April 2020.
**Annex 1: A sample of the projects that have been completed via the LIDA Data Scientist Internship Programme**

<table>
<thead>
<tr>
<th>Project title</th>
<th>Principal Investigator</th>
<th>Institute/School</th>
<th>Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td>A data driven approach to understanding blood cancers</td>
<td>Prof. Dave Westhead</td>
<td>Molecular and Cellular Biology</td>
<td>FBS</td>
</tr>
<tr>
<td>Investigating factors behind low or zero household recycling rates in Leeds</td>
<td>Prof. William Young</td>
<td>Earth and Environment</td>
<td>ENV</td>
</tr>
<tr>
<td>Synthetic Population Estimation and Scenario Projection – The English Future Elderly Model</td>
<td>Dr Nik Lomax</td>
<td>Geography</td>
<td>ENV</td>
</tr>
<tr>
<td>Application of natural language processing for identification of online hate on Twitter</td>
<td>Dr Carly Lightowlers</td>
<td>Law</td>
<td>ESSL</td>
</tr>
<tr>
<td>Petition Parliament Debates on Twitter: Assessing the effectiveness of the e-petition procedure through Twitter conversations</td>
<td>Dr Viktoria Spaiser</td>
<td>POLIS</td>
<td>ESSL</td>
</tr>
<tr>
<td>An area classification of consumer vulnerability in the UK</td>
<td>Dr Nik Lomax</td>
<td>Geography</td>
<td>ENV</td>
</tr>
<tr>
<td>Multidisciplinary working across perioperative medicine and primary care: a health informatics feasibility study</td>
<td>Dr Simon Howell</td>
<td>LIHS</td>
<td>FMH</td>
</tr>
<tr>
<td>Textile Data Analytics (TDA): Does fabric tactility affect clothing product sales?</td>
<td>Dr Ningtao Mao</td>
<td>Design</td>
<td>FAHC</td>
</tr>
<tr>
<td>Developing personalised nutrition: linking diet and health data</td>
<td>Prof. Janet Cade</td>
<td>Mathematics</td>
<td>MAPS</td>
</tr>
<tr>
<td>Understanding customers’ missions from the products they purchase</td>
<td>Prof. Roy Ruddle</td>
<td>Computing</td>
<td>ENG</td>
</tr>
<tr>
<td>Extracting actionable insights from free text police data</td>
<td>Dr Daniel Birks</td>
<td>Law</td>
<td>ESSL</td>
</tr>
<tr>
<td>A machine learning approach to understanding the disease trajectories of atrial fibrillation</td>
<td>Dr Jianhua Wu</td>
<td>Dentistry</td>
<td>FMH</td>
</tr>
<tr>
<td>Predicting and warning extreme wind response in bridges using advanced data analytics</td>
<td>Drs Antonio Abellan Fernandez &amp; Nikolaos Nikitas</td>
<td>Civil Engineering/Geography</td>
<td>ENG &amp; ENV</td>
</tr>
<tr>
<td>Probabilistic Programming and Data Assimilation for Next Generation City Simulation</td>
<td>Prof Nick Malleson &amp; Dr Jonathan Ward</td>
<td>Geography &amp; Mathematics</td>
<td>ENV &amp; MAPS</td>
</tr>
<tr>
<td>Classification of primary brain tumours based on their predicted response to standard therapy</td>
<td>Dr Lucy Stead &amp; Prof Dave Westhead</td>
<td>Molecular and Cellular Biology</td>
<td>FBS</td>
</tr>
<tr>
<td>Can robots learn to drive like humans? Using inverse reinforcement learning to imitate human driving in autonomous vehicles</td>
<td>Dr Minh Le Kieu &amp; Profs Alison Heppenstall &amp; Ed Manley</td>
<td>Geography</td>
<td>ENV</td>
</tr>
<tr>
<td>Developing an algorithm to identify cryptic splice variants in whole exome datasets</td>
<td>Dr James Poulter &amp; Prof Colin Johnson</td>
<td>LIRMM</td>
<td>FMH</td>
</tr>
<tr>
<td>Translation Difficulty Estimation</td>
<td>Prof Serge Sharoff</td>
<td>Languages, Cultures &amp; Societies</td>
<td>LCS</td>
</tr>
</tbody>
</table>
Annex 2: LIDA Data Scientist Internship Project Proposal

Project title:
Primary Academic supervisor:
Any other supervisors:
School(s)/Institute(s):
External partner(s):
Funding type: Self-funding / LIDA funding requested
N.B. if self-funding, please provide the contact name of Faculty Research Office colleague who will be setting up the project account:
Preferred start date and duration (delete as appropriate): Monday 5th October 2020 or Monday 5th April 2021

1. Underpinning research

Provide a summary of the research which underpins the proposed work, including project/grant codes where possible (max. 300 words).

2. Project summary

Provide a summary of the project and its main aims and objectives (max. 300 words).

This should be written in the form of a project abstract for a broad, external audience as this may be used in literature about the Programme circulated with other project supervisory teams. Please indicate if any information is sensitive and not for general disclosure.

3. External partner(s)

Give brief details of the following:
i) An overview of the company/external partner and outline the current relationship with them. (Max 100 words.)
ii) The resources the external partner will contribute to the project (Max 200 words.).
iii) Whether an NDA is in place with the partner.

A letter of support from the external partner(s) also detailing this level of cash and in-kind

---

1 Please indicate the status of this funding (i.e. is it secured or only bid for at this stage?) and when you expect to hear if you have secured funding. Please note monies used for self-funded projects should not be those granted for the purposes of employing research staff due to the fact that the interns are employed on technical contracts and do not qualify for this funding.
contribution must be submitted alongside this application.

4. Data source(s), volume and sensitivity

Provide details of the data source(s) that will be used during the course of the project. Please note that you should already have access to the proposed data or be able to provide details that access is on track ready for the start of the project.

Please specify:

- any data sharing agreements that are already in place with third parties;
- details of the sensitivity of the data;
- how the data is currently accessed and stored;
- whether you will require any support from the LIDA Data Analytics Team.

Please also indicate whether your project will require data held by the CDRC as detailed in its Data Store. (max. 300 words).

Please note – proposals which already have access to their data, or where the data is aggregated and does not require use of a safe room (IRC secure) will be seen as stronger.

5. Methods and Software

Provide a summary of the methods the intern will be using and any software packages, computer storage or HPC capability they will need to be provided with in order to carry out the work. (max. 300 words)

6. Outputs, outcomes and future plans

Provide a summary of the output(s) and outcome(s) that will arise from the project and future plans relating to this work (max. 300 words).

It is important that you specify the ways in which the intern will be involved in these outputs/outcomes.
7. Essential skills and requirements

Provide details of any specific skills the intern should possess in order to complete the proposed project, e.g. familiarity with a particular programming language. (max. 300 words).

Please note the essential and desirable criteria under which the interns will be recruited on p.2 above and bear this in mind for this section.

For further information about the LIDA Data Scientist Internship Programme please contact:

Kylie Norman, Administrator for the Programme: k.r.norman@leeds.ac.uk
Or Paul Evans, Business Development Manager: p.d.evans@leeds.ac.uk

Further information about the Data Scientist Internship Programme can be found at: https://lida.leeds.ac.uk/study-training/data-science-internship-scheme/