We would like to solicit comments on the general structure for the training year of the data science CDT bid. We are tabling this to get feedback on the general idea that we can incorporate into a subsequent full proposal to Board of Studies.

**Type of Programme**

Taught MRes

**Sketch of Learning Objectives**

By “data science” we mean computational methods that have the goal of extracting knowledge from large-scale corpora of unstructured data like text, images, speech, video, etc. This is an interdisciplinary combination of machine learning, natural language processing, databases, algorithms (e.g., streaming and randomized algorithms), speech processing, and computer vision. The goal of the CDT is to prepare students to perform original research in these areas. To support this, the learning objectives of the MSc year will be:

- *Breadth of knowledge* across the data sciences
- *Advanced technical background* in at least one of the data science areas
- *Overview of current research problems* across the data sciences
- *Appreciation for real-world problems* involving the use of data in industry, science, and the public sector
- *Research experience* in one of the data sciences

In addition it is a key goal of EPSRC that the students each year form a “cohort”, which we interpret to mean a cohesive group with a distinctive identity that seek out opportunities to learn from each other. This seems no bad thing.

**Degree Structure**

The degree will be a taught MSc by Research, designed to allow continuous registration into the PhD using a 1+3 scheme. The degree will consist of

- 60 points level 11 courses drawn from existing provision
- 30 point new course “Introduction to Research in the Data Sciences”
- 90 point project. This would take up 50% of the spring and all of the summer.

Graphically, the degree structure is

Other structures were considered, but they seemed less well suited to our learning objectives:
1. A programme along the lines of our current taught MSc would have only a 60 pt project. A larger scale project would allow students a better introduction to research methods, including the possibility of producing publishable research in their first year. Even the very best MSc students seldom do this in our current taught degree.

2. On the other hand, if the project were any larger, e.g., 120pts or 140pts, then students would need to be assigned a project and a supervisor as soon as they arrive. We anticipate that this would have the effect of causing students to align more with their MSc supervisor than to the CDT, reducing the cohort effect. In our view, the Neuroinformatics DTC has done an outstanding job of creating a group of students that have camaraderie and learn from each other. We suspect that part of the reason for this is that students are not immediately matched with a supervisor until they have gained a bit of background and experience. Furthermore, more than 40pts of existing courses are required to achieve both the goals of breadth and advanced technical knowledge in a specific area.

3. At the other extreme, one could imagine a 180pt MRes or even a 4 year PhD programme without an MSc component (NB CDT bids in other schools are envisioning this). The difficulty with this is that a taught component is both required by EPSRC, and desirable in terms of the learning objectives above. Without the 1+3 structure, it is difficult to see how we would require students to pass the taught component in order to obtain the PhD.

Existing Informatics Courses

We will draw extensively from existing Informatics courses. This will be subject to a breadth requirement, i.e., that students must take at least one course each in: (a) Machine Learning, (b) “Data Management” (Databases, EXC), and (c) Applications (e.g., NLP, vision, speech). The courses that we would provisionally consider under each of these headings are:

<table>
<thead>
<tr>
<th>Machine Learning</th>
<th>Data Management</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMR</td>
<td>DIE</td>
<td>FNLP</td>
</tr>
<tr>
<td>MLPR</td>
<td>ADBS</td>
<td>ANLP</td>
</tr>
<tr>
<td>IAML</td>
<td>QSX</td>
<td>BIO1</td>
</tr>
<tr>
<td>IT</td>
<td>TDD</td>
<td>BIO2</td>
</tr>
<tr>
<td>RL</td>
<td>DBS</td>
<td>TTS</td>
</tr>
<tr>
<td></td>
<td>EXC</td>
<td>MT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NLG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NIP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLSC</td>
</tr>
</tbody>
</table>

New Courses

We envision a new core course to the CDT, called “Introduction to Research in Data Science,” which would be required of all CDT students. In addition, we imagine “topics” or seminar courses to give students more of an overview of current literature.
Two different models for this are described for comment below: namely, the *Independent Research Review in Data Science*, and the *Topics in Data Science* courses. It is unlikely that we would run both courses, so comments about their feasibility would be useful. All courses will be open only to CDT students, and also to all PhD students in the school.

**Introduction to Research in Data Science (required)**

The goals of this course are twofold: to give students a broad overview of the current research landscape in data science, and to give students an appreciation for practical data science problems in industry, other academic disciplines, and the public sector.

The first component of the course consists of guest lectures by the PhD supervisors in the CDT. Each will give an overview of the research landscape in their area, including research currently being done in their group.

The second component of the course will be a “small practical project.” This is intended to be a small project that combines: (a) ideas from one of the taught courses the student is taking, (b) a topic of current research interest, and (c) a problem or data set from the “real world.” The CDT will solicit data sets and problems from external partners in order to support this project. This scheme is patterned off of the “class project” which is common in PhD-level courses in North America. This project has a different objective from the larger MRes dissertation, because a student might legitimately choose to do a more theoretical topic for the MRes and PhD projects, while the small practical is designed to have a more practical component.

The small practical project will be assisted by TAs who are more senior CDT students, using the model from IRR/IRP. We hope that this will enhance the cohort effect, provide better induction for first years, and provide opportunities for peer-to-peer learning. We expect a 1 to 3 ratio of tutors to first years. The tutors will be funded by the CDT.

We propose that this course be pass / fail. It will be evaluated by (a) a report based on the class project (12-15 sides), and (b) a poster presentation based on the project. We will have a poster session at the end of the course to which all members of Informatics will be invited. The course lecturers will evaluate the quality of the students’ poster presentations and provide constructive feedback where appropriate.

**Independent Research Review in Data Science (optional)**

This would be a new 10pt course in which each student does in-depth reading of the current literature in a chosen topic. Delivery is based on a series of 1:1 meetings with a member of the CDT academic staff. The student and the staff member should agree on a coherent series of readings on a specialized research topic. The student produces two documents: a proposal by the end of Week 4 that gives a short outline of the area (roughly a side) plus an expected bibliography, and at the end of the semester, a term paper describing the previous work. The review is expected to be a critical review, comparing and contrasting related papers, discussing shortcomings of the existing literature, and in some cases proposing follow-on research.

Because this course is based on a series of 1:1 meetings with a mentor, it can be flexibly timetabled, i.e., students can take it either in the autumn or spring.

**Topics in Data Science (optional)**

This would be a small seminar course (10pts) in which students read, discuss, and critically review recent research papers in group setting. Topics will change from year to year based on changing trends in the research community, as well as the interests of the students as the lecturer. Existing PhD students would be welcome to participate as well. Evaluation would be by: (a) a series of written responses to the papers each week, and (b) either an exam on the readings, or a short term paper. This course could be resourced by research staff as well as academic staff.

**Application Arrangements**

We will construct an admissions committee for the CDT that reviews applications, interviews candidates, and makes admissions decisions. This will be patterned off of the Neuroinformatics DTC.

**Students with previous MSc**

The CDT committee will examine the students’ record to assess the degree of overlap with existing courses. Students whose overlap with our existing courses is small will be encouraged to take the standard MSc in Data Sciences. Students with substantial overlap (e.g., previous MInf here with many courses in ML and databases) can with the approval of the CDT training director be registered directly into the PhD programme with a four year prescribed programme of study. Such students would be expected to audit the Introduction to Data Sciences course.
IAD Resources

The PhD programme as a whole will include an IAD component, peer-to-peer learning, for training in research skills, writing and presentation skills, leadership skills, and entrepreneurship / business skills. However we will focus these in years 2-4. With 6 taught courses, students are likely to be too busy to take this component seriously. These will include seminars taught from the IAD, industry events such as those led by Informatics Ventures, and panel discussions led by staff, higher-year PhD students, and PhD alumnae.