**PhD Project Proposal**

School of Electronics, Electrical Engineering and Computer Science

& ECIT Global Research Institute

|  |
| --- |
| **Proposed Project Title: Machine learning for heterogeneous data integration** |
| **Principal Supervisor:** DrAnna Jurek-Loughrey |
| **Project Description:**  With the increasing volume of information available from various sources, data integration has become an emerging topic in the area of big data analytics. Integrating data from multiple sources drastically expands the power of information and allows us to address questions that are impossible to answer using a single data source. For example, integrating records from law enforcement watch lists with data coming from the web could help to prevent terrorist attacks. As a part of the data integration process, records that refer to the same real world entity (e.g. person) need to be identified and linked. This task, referred to as entity resolution, is challenging in particular when the data is presented in diverse and unstructured formats.  The main focus of this project will be to develop entity resolution methods for heterogeneous data integration with the following specific objectives:   * Review and understand the limitations of the current state-of-the-art research in the area * Explore metrics for measuring similarities between mixed data formats (e.g. structured and unstructured text) * Develop algorithms for identifying connections amongst data records originating from independent sources (e.g. records from different datasets that refer to the same person) * Evaluate the proposed algorithms with complex real world datasets |
| **Contact details**  Supervisor Name: Anna Jurek-Loughrey Tel: +44 (0)28 9097 4484  QUB Address:  Computer Science Building  03.009, 18 Malone Road Email: a.jurek@qub.ac.uk  BT9 5BN Belfast |