UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Qualifying Exam

Tuesday October 31, 2023 at 10AM, In-Person REYN 1101

Jeremy Foxcroft

Reconstructing Canadian Identities:
Techniques and Strategies for Historical Record Linkage

Chair: Dr. Neil Bruce
Advisor: Dr. Luiza Antonie
Co-Advisor: Dr. Kris Inwood [Economics]
Non-Advisory 1: Dr. Fattane Zarrinkalam [SoE]
Non-Advisory 2: Dr. Andrew Hamilton-Wright

Abstract:

Record linkage is the process by which two data records, often from different data sources, are identified as referring to the same entity. A common example might be two collections of records that refer to people (e.g., prison records, hospital records, census records, birth/death/marriage records, etc.). When records can be uniquely identified by a common attribute (e.g., a social security number) this task is trivial. When records do not have a unique common attribute, common attributes they do share (e.g., name, date of birth, etc.) can be used to identify similar records and classify them as matching or non-matching. Economists, historians, and demographers can use the resulting linked population data to investigate previously unanswerable questions.

This proposal discusses performing record linkage on historical data, primarily Canadian census records from the 19th and 20th centuries. What data was collected (e.g., what are the shared questions between different censuses), how it was collected (e.g., how are non-English names anglicized), and when it was collected (e.g., have people relocated since the last census) all make this task uniquely challenging.

Using a limited number of records from different census collections which are known to refer to the same individuals, supervised machine learning classification models will be used to link entire census populations. Different linkage methodologies will yield different sets of links, all of which will be analysed for various forms of bias.

Given names and surnames are two of the most important identifiers consistently available between years, and this proposal also outlines new string similarity measures targeting these attributes. In addition, this work explores the training data construction process by examining how the class ratio in training data affects the most commonly used performance measures for record linkage.