36-491/691 DATA MATCHING METHODS AND THEIR USES
6 Units, MINI 1, Fall 2013

Instructors: Stephen E. Fienberg, Rebecca Nugent
Baker Hall 132G, 232C
(412) 268-2723/7830
fienberg, rnugent@stat.cmu.edu
http://www.stat.cmu.edu/∼fienberg, ∼rnugent
Office Hours: TBA

TAs: Sam Ventura
sventura@stat.cmu.edu
Office Hours: TBA

Class Meetings: Monday and Wednesday, 2pm-3:20pm, Baker 232M

Website: http://www.cmu.edu/blackboard,http://www.stat.cmu.edu/NCRN

Pre/Co-requisites: Computing at the level of 36-350, Data Analysis at the level of 36-401, *Instructor Permission Required*

Textbooks: Both textbooks available through the class.
Data Quality and Record Linkage Techniques by Herzog, Scheuren, Winkler, Springer (2007)
Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection by Peter Christen, Springer (2012).

General Course Plan: This course is an introduction to "data matching", the field of improving data quality through record linkage techniques including merging lists, identifying duplicates, and unique entity resolution. While always an active area of interest, the increasing size and use of databases in today's large-scale data analysis and modeling problems makes this topic more relevant than ever. The results of a model or statistical analysis are always dependent on the data; these methods attempt to create the "correct" data set. For example, hospital databases might have two records for "Rebecca Nugent' and "Becca Nugetn"; knowing that they refer to the same person might be crucial for medical and/or insurance purposes. Or there might be credit check information for both "Steve Fienberg" and "Stephen Feinberg", a fact that may cause issues when trying to get a mortgage. Or people may be receiving benefits or payouts illegally under multiple names - how do we find them? Even correctly identifying homicide victims in Columbia or casualties of the current conflict in Syria require record linkage methods. Typographical errors, missing information, and perhaps politically-motivated differing lists all contribute to the need for statistical matching algorithms. These algorithms obviously need to match records correctly with a high probability, but also should scale to very large databases. Imagine trying to match records for all 315 million people in the United States - a task being carried out by the U.S. Census Bureau every day. In this class, we will discuss working with text and its similarity metrics, imputation models for missing data, statistical models for predicting matches, and real-life case studies of record linkage applications.

Course Objectives:

1. Learn when data matching methods are needed and how they are used
2. Develop skills to effectively work with and analyze text data
3. Identify and implement record linkage techniques appropriate to the application's goal
4. Develop written and verbal communication skills for discussing conclusions and limitations of statistical evidence; present data analysis appropriately in a scientific report
5. Effectively use R, a widely-used statistical package, in data analysis

Course Work: Your grade in this course will be determined by homework assignments, a final project, and discussion participation.

- Weekly homework assignments are due at the beginning of class on Wednesdays. The type of homework submission (paper vs electronic) will be dependent on what is assigned. Deviation from the requested format requires instructor permission.

  Please see the TA or instructor during office hours for help with homework problems. Questions posed by email must be sent at least 24 hours before the time an assignment is due in order to guarantee a response.

- The final project will be discussed later. A presentation and report are required.

- This course will be taught with the flavor of a discussion seminar or research working group. Active participation is required and encouraged. Students should feel comfortable offering their thoughts and ideas.

Grading policy: You may discuss homework problems with your fellow students, however the work you submit must be your own. Acknowledge any help received on your assignments. Copied assignments will receive no credit. Late assignments require instructor permission.

You have one week from the day an assignment is handed back in class to bring any grading issues, comments, complaints, etc to the attention of the instructor. Please note that if you are absent the day something is handed back, this deadline will not be extended unless arrangements have been made in advance with the instructor.

Final grades will be computed with the following weights:

Homework	.60
Final Project	.30
Participation	.10

Final letter grades will be determined as usual: [90,100] = A, [80,89] = B, [70,79] = C, [60,69] = D, [$<$ 60] = R. Grades may be curved at the instructor's discretion (effort, improvement, etc).

Computing: The statistical computing package we will use in this course is R. R is available on many campus computers, and you may download a free version from `www.r-project.org`. You may also use the nearly-identical (but not free) program called S+, available on all campus computers. You can obtain a free temporary version from `myandrew`. This version is good for 1 year; you can keep renewing the license as long as you are a CMU student.

R References: manuals available on R website;
`http://www.stat.cmu.edu/~rnugent/teaching/introR`
Introductory Statistics with R, Peter Dalgaard; Springer-Verlag
Modern Applied Statistics with S-Plus Venables, Ripley; Springer

Laptop Policy: Students are expected to be participating in class; any laptop use during class should pertain directly to the class. Instructor reserves the right to not allow laptop use during class. When the class has a guest speaker, laptops must be turned off and put away.

Cellphones/Pagers, etc: All cellphones, pagers, beepers, and anything else that makes noise should either be turned off or silenced during class.

Communication: Assignments and class information will be posted on Blackboard. Help with using blackboard is available at www.cmu.edu/blackboard/help/.

Email: Sending email to your professor or teaching assistants should be treated as professional communication. Emails should have an appropriate greeting and ending; students should refrain from using any kind of "shortcuts", abbreviations, acronyms, slang, etc. in the email text. Emails not meeting these standards may not be answered.

Academic Integrity: All students are expected to comply with the CMU policy on academic integrity. This policy is online at www.studentaffairs.cmu.edu/acad_integ/acad_int.html

Cheating, copying, etc will not be tolerated; please ask if you unsure of whether or not your actions are complying with assignment/exam instructions. Always ask if you are unsure; always default to acknowledging any help received.

Video/Audiotaping: No student may record or tape any classroom activity without the express written consent of the professor. If a student believes that he/she is disabled and needs to record or tape classroom activities, he/she should contact the Office of Equal Opportunity Services, Disability Resources to request an appropriate accommodation.

Disability Services: If you have a disability and need special accomodations in this class, please contact the professor. You may also want to contact the Disability Resources office at 8-2013.

TENTATIVE SCHEDULE *subject to change*

Date	**Topic**	**Due**
Mon 8/26	no class	
Wed 8/28	Overview of Record Linkage	
Mon 9/2	no class (Labor Day)	
Wed 9/4	Data Collection; Case Studies	HW 1
Mon 9/9	Text Similarity Metrics; Supervised Models	
Wed 9/11	ROC Curves, Deterministic vs Probabilistic	
Mon 9/16	Logistic Regression; Classification Trees	
Wed 9/18	Fellegi-Sunter Model	HW 2
Fri 9/20	**Special Class: Conflict in Syria**	
Mon 9/23	Fellegi-Sunter Model	
Wed 9/25	E-M Algorithm; Using Logistic Regression with F-S	HW 3 (9/27)
Mon 9/30	Blocking; Capture/Recapture	
Wed 10/2	Transitivity, Clustering Records; Matching more than 2 lists	
Mon 10/7	Record Linkage with too much data	
Wed 10/9	Final Projects	
Mon 10/14	Final Projects	