STA 794: Modern Advancements in Record Linkage
Duke University, Fall 2016

Instructor: Rebecca C. Steorts, Assistant Professor,
Department of Statistical Science and Computer Science, beka@stat.duke.edu, Old Chem 216
Course Time: Tuesday: 10:00AM – 1:00 PM
Steorts Office Hours: Tuesday: 1:30 – 2:30 PM, Old Chem 216
Course webpage: `https://stat.duke.edu/~rcs46/linkage.html`

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. Among the types of questions that have been, and can be, addressed by combining information include: How accurate are census enumerations for minority groups [22, 23]? How many of the elderly are at high risk for sepsis in different parts of the country [19]? How many people were victims of war crimes in recent conflicts in Syria [17]?

In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts need to use methods from statistical and computational science known as *record linkage* (also called *entity resolution* or *de-duplication*) to proceed with analysis. Record linkage is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself, because many databases contain errors (noise, lies, omissions, duplications, etc.), and the number of parameters to be estimated grows with the number of records [1–16, 18, 20, 21].

The objective of this course are to provide an introduction to record linkage methodology and computational tools. This will be achieved by reading papers, lectures, group discussions, and student led discussions on papers that are assigned weekly. Students will have the opportunity to complete computational (coding) tasks to better their understanding of record linkage and how it is useful on both synthetic and real data sets.

Prerequisites The course is appropriate for graduate students in statistics, electrical engineering, computer science, mathematics, and related fields who have a strong background in applied statistics and Bayesian methods. Students are expected to be very familiar with `R` and are **encouraged** to have learned `LaTex` by the end of the course. All code you write should be reproducible in R markdown. (Other software environments are completely fine; just please check with Professor Steorts about these).

Recommended Textbook: *Data Matching*, Peter Christan, 2012, `http://www.springer.com/gp/book/9783642311635`.

Homework: Homework will be assigned for each class period in the form of reading a particular paper and submitting a summary of that paper. Each student should prepare slides for their presentation, which should be turned in as well. Summaries should be a couple of paragraphs each, no more than a page. Homework should be done individually, though students are encouraged to

help each other. There will always be at least one week's notice on each paper. There will also be coding tasks. These will be posted on the course webpage. There will be 6 total coding tasks for the semester. You need to complete 3 of the 6 coding tasks.

Presentations: Your presentations for each paper you present should be approximately 30 minutes long. You should do the following:

1. Outline the main ideas of the paper and give a motivating example.
2. Outline the methods and the ideas.
3. Outline any algorithms.
4. Do you want to show how the methods/algorithms work in practice on very simple data. (You might also share your code with the class and do a class demo).
5. You could also illustrate the methods on real data.
6. Please have a discussion section for the class listing pros and cons of the methods and applications in the paper.

Please be sure to turn your slides and all materials by 11:59 PM the day before you present the material. Please send these via email to Professor Steorts.

Grading Policy: Evaluation will be based on the presentation of papers/summaries and the coding tasks. If you have a question regarding how you are doing in the class, feel free to set up a time to meet with me. Assignments will be posted on the course webpage. You are to submit your assignments through Sakai by the due date and time. Grades will not be calculated on Sakai.

Presenting papers, i.e. slides (3 for the semester):	25%
Summaries of papers (3 for the semester):	25%
Coding tasks (3 for the semester):	50%

An overall score of s will result in a grade of Satisfactory/Unsatisfactory (Pass/Fail) basis:

S if $70 \leq s \leq 100$
U if $0 \leq s < 60$.

Course Policies: All coding assignments will be announced in class along with the due date. You must complete 3 of the 6 total coding tasks over the course of the semester, with a passing grade. Coding tasks must be turned in electronically through Sakai. **Late coding tasks will not be accepted**.

Academic Honesty: Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. Cheating on exams and quizzes,

plagiarism on homework assignments, projects, and code, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved as well as being reported to the University Judicial Board. Additionally, there may be penalties to your final class grade. Please review Duke's Standards of Conduct. For more information on the Duke honor code (known as Duke Community Standard), please go to `http://integrity.duke.edu/faq/faq1.html`.

Students with Disabilities: Students who require special accommodations in class or during exams should follow the procedures outlined by the Disability Management Program http://access.duke.edu/students. Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made.

Privacy Policies: Student records are confidential.

References

[1] Ball, P. (2000). The Salvadoran Human Rights Commission: Data Processing, Data Representation, and Generating Analytical Reports. In *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis* (P. Ball, H. F. Spirer and L. Spirer, eds.). AAAS.

[2] Bhattacharya, I. and Getoor, L. (2006). A latent dirichlet model for unsupervised entity resolution. In *SDM*, vol. 5. SIAM.

[3] Bilenko, M. and Mooney, R. J. (2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *KDD '03*. ACM, 39–48.

[4] Christen, P. (2008). Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification. In *KDD '08*. ACM, 151–159.

[5] Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, **24** 1537–1555.

[6] Cohen, W., Ravikumar, P. and Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, vol. 3. 73–78.

[7] Dai, A. M. and Storkey, A. J. (2011). The grouped author-topic model for unsupervised entity resolution. In *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 241–249.

[8] Gutman, R., Afendulis, C. and Zaslavsky, A. (2013). A bayesian procedure for file linking to analyze end- of-life medical costs. *Journal of the American Statistical Association*, **108** 34–47.

[9] Hsu, W., Lee, M. L., Liu, B. and Ling, T. W. (2000). Exploration Mining in Diabetic Patients Databases: Findings and Conclusions. In *KDD '00*. ACM, 430–436.

[10] Jewell, N. P., Spagat, M. and Jewell, B. L. (2013). MSE and Casualty Counts: Assumptions, Interpretation, and Challenges. In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict* (T. B. Seybolt, J. D. Aronson and B. Fischhoff, eds.). Oxford University Press, Oxford, UK.

[11] Larsen, M. D. (2002). Comments on Hierarchical Bayesian Record Linkage. In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. The American Statistical Association, 1995–2000.

[12] Larsen, M. D. (2005). Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory. In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. The American Statistical Association, 3277–3284.

[13] Lum, K., Price, M. E. and Banks, D. (2013). Applications of Multiple Systems Estimation in Human Rights Research. *The American Statistician*, **67** 191–200.

[14] McCallum, A. and Wellner, B. (2004). Conditional Models of Identity Uncertainty with Application to Noun Coreference. In *Advances in Neural Information Processing Systems (NIPS '04)*. MIT Press, 905–912.

[15] Miller, P. L., Frawley, S. J. and Sayward, F. G. (2000). IMM/Scrub: A Domain-Specific Tool for the Deduplication of Vaccination History Records in Childhood Immunization Registries. *Computers and Biomedical Research*, **33** 126–143.

[16] Murphy, J., Brackbill, R. M., Thalji, L., Dolan, M., Pulliam, P. and Walker, D. J. (2007). Measuring and Maximizing Coverage in the World Trade Center Health Registry. *Statistics in Medicine*, **26** 1688–1701.

[17] Price, M., Klinger, J., Qtiesh, A. and Ball, P. (2013). Full updated statistical analysis of documentation of killing in the Syrian Arab Repulic. Human Rights Data Analysis Group, commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR).

[18] Sadinle, M. (2014). Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach. *Annals of Applied Statistics*, **8** 2404–2434.

[19] Saria, S. (2014). A 3 trillion challenge to computational scientists: Transforming healthcare delivery. *IEEE Intelligent Systems*, **29** 82–87.

[20] Sariyar, M. and Borg, A. (2010). The RecordLinkage Package: Detecting Errors in Data. *The R Journal*, **2** 61–67.

[21] Sariyar, M., Borg, A. and Pommerening, K. (2012). Active Learning Strategies for the Deduplication of Electronic Patient Data Using Classification Trees. *Journal of Biomedical Informatics*, **45** 893–900.

[22] Seybolt, T. B., Aronson, J. D. and Fischhoff, B. (2013). *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*. Oxford University Press.

[23] Winkler, W. E. (2006). Overview of record linkage and current research directions. In *Bureau of the Census*. Citeseer.