STA 490/690: Almost All of Entity Resolution
Duke University, Spring 2021

Please note that the syllabus may change if needed and with the agreement of the majority of the students. We will adapt to online learning in a way which is most suited for the entire course, which will be measured by surveys by students.

1 Course Schedule

Instructor: Rebecca C. Steorts, Assistant Professor, Department of Statistical Science
Email: beka@stat.duke.edu
Course Time: Wednesday/Friday: Noon – 1:15 PM EDT
Office Hour: TBD
Revised academic calendar: https://registrar.duke.edu/fall-2020-academic-calendar

2 Course Description

Information about social entities is often 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? How many of the elderly are at high risk for sepsis in different parts of the country? How many people were victims of war crimes in recent conflicts in Syria?

In most tasks, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because identifiers don't exist in some databases or are not available due to privacy concerns. Entity resolution, an important tool in statistics, machine learning, and computer science, is used to remove the duplications often found in the absence of a unique identifier. In this course, I will provide an overview of common methods and algorithms that are used every day in industry by top tech and finance companies to merge data and make predictions. This is a great course to take if you're interested in applying for internships or jobs in industry or pursuing an honor's thesis/research!

The course can count toward the undergraduate data science concentration. This course can also count towards the undergraduate statistical science major.

Pre-reqs: STA 360 or permission of the instructor

3 Course Learning Objectives

In this course, the following learning objectives will be met by the end of the semester:

1. Learn machine learning, statistical, and computer science methods and how they related to industry/academia
2. Learn to work in a reproducible environment using GitHub
3. Learn the core entity resolution methods and algorithms that are used in industry
4. Work on case studies for entity resolution using real data sets in groups or solo
5. Work on a small case study at the end of the semester in groups/solo. Goals include preparing a small written report (4 pages), short presentation (5 minutes), reproducible code, and evaluation of other projects. This is meant to be a longer homework assignment that you will have time to work on in class and work on with your groups. The goal of this project is that you can use this project and talk about it orally when applying for jobs in your portfolio and on on your resume.
6. Learning many different software packages via demos by the instructor and reinforcing these on short interactive exercises in class or homework exercises.
7. Getting exposure to distributed computing via Apache Spark and understanding fundamentals of connections between statistical science and machine learning

4 Course Community

Duke Community Standard As a student in this course, you have agreed to uphold the Duke Community Standard, which can be found at `https://studentaffairs.duke.edu/conduct/about-us/duke-community-standard`. You also have agreed to the practices specific to this course. A video regarding the Duke Community Standard can be found here at `https://youtu.be/_KN97j30ST4`.

Inclusive Community It is my intent that students from all diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that the students bring to this class be viewed as a resource, strength and benefit. It is my intent to present materials and activities that are respectful of diversity and in alignment with Duke's Commitment to Diversity and Inclusion. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.

Furthermore, I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities. To help accomplish this:

1. If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. If you prefer to speak with someone outside of the course, your academic dean is an excellent resource.

2. I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to me about it.

Accessibility If there is any portion of the course that is not accessible to you due to challenges with technology or the course format, please let me know so we can make accommodations.

In addition to accessibility issues experienced during the typical academic year, I recognize that remote learning may present additional challenges. Students may be experiencing unreliable wi-fi, lack of access to quiet study spaces, varied time-zones, or additional responsibilities while studying at home. If you are experiencing these or other difficulties, please contact me at the earliest opportunity to discuss possible accommodations.

The Student Disability Access Office (SDAO) is available to ensure that students are able to engage with their courses and related assignments. Students should be in touch with the Student Disability Access Office to request or update accommodations under these circumstances.

Academic honesty Don't cheat!
Please abide by the following as you work on assignments in this course:

1. You may not discuss or otherwise work with others on the exams. Unauthorized collaboration or using unauthorized materials will be considered a violation for all students involved. More details will be given closer to the exam date.

2. Reusing code: Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. If you directly use code from an outside source (or use it as inspiration), you must explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.

3. On individual assignments, you may not directly share code or write up with other students. On team assignments, you may not directly share code or write up with another team. Unauthorized sharing of the code or write up will be considered a violation for all students involved.

Any violations in academic honesty standards as outlined in the Duke Community Standard and those specific to this course will automatically result in a 0 for the assignment and may result in a failing grade for the course. Violations will be reported to the Office of Student Conduct for further action.

5 Course Activities

Communication All lecture notes, assignment instructions, up-to-date schedule, and other course materials may be found on the course website: **update when ready**. The course webpage (including the course github page) will be updated regularly, so please make sure to check for updates often.

Announcements will be made during class or during Piazza, so please check your email regularly. I will send out surveys periodically to gauge how the class is going, so please respond to these as I greatly appreciate your feedback to improve the course!

Teaching Team There is an excellent team of teaching assistants in place. They are here to help and assist you throughout the course!

1. Teaching Assistant, Email

Lecture Component Lectures will have two components:

1. **Lecture content videos**: These are pre-recorded videos that contain content of the lecture, and can be treated as a "video textbook." You should watch these before coming to the live session. **If you are having trouble viewing the videos, please install** `https://www.videolan.org/vlc/download-windows.html`**.**
2. **Live lecture sessions**: These sessions will be on Zoom during the scheduled class period. The majority of class will be dedicated to the following:
   (a) Prof. Steorts will go through the concepts in class again, highlighting the most important parts, providing clarifications, solutions/advice, and more advanced insights.
   (b) Prof. Steorts will take questions (in person) or those that are emailed to her by Monday at 10 AM EDT and Wed at 10 AM EDT.
   (c) Prof. Steorts will provide exercises for the class to work through, practice exams, and an interactive environment.
   (d) TAs will be present to help in these activities such that we can break into small groups and students can receive more individualized attention.

Getting Help If you have a question during lecture or lab, feel free to ask it! There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.

The teaching team is here to help you be successful in the course. You are encouraged to attend office hours during the times posted on the home page to ask questions about the course content and assignments. A lot of questions are most effectively answered *remotely via video*, so office hours are a valuable resource. Please use them!

Outside of class and office hours, any general questions about course content or assignments should be posted on Piazza. You can access the group from the course home page or via Sakai.

Please find the following schedule regarding who is responsible for answering the Pizza on a particular day.

6 Prior Knowledge, Course Expectations, and Grading Policies

Prior Knowledge Students are expected to have a solid background in regression analysis (STA210), elementary probability (STA 230 or STA 240), and elementary linear algebra (MATH 202/216/218/221), and STA 360/601/602. These will be building blocks for course topics, and very little review will be provided in this course. If you are unsure what prior knowledge is expected, please refer to the following past syllabi & resources, and review any gaps in knowledge before the start of the semester.

1. STA 210: `https://www2.stat.duke.edu/courses/Spring19/sta210.001/`
2. STA 230: `https://www2.stat.duke.edu/courses/Fall18/sta230/`
3. Linear algebra: `http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/lecture_12`
4. R programming. Students are expected to have a solid foundation of R programming prior to the first day of lecture. For a review of R programming, please see the following **review lectures and videos** on R at `https://github.com/resteorts/modern-bayes`.
5. github: I expect that you have a solid foundation of using github prior to the first day of class as this will be important for accessing all class resources. Resources can be found at `https://github.com/resteorts/modern-bayes`
6. Common Distributions. I will assume that you have a good handle on common distributions from your probability class. In case you need to review this, there is a good reference, here: `https://github.com/resteorts/modern-bayes/blob/master/reading/statistical-inference.pdf`, where you can review common probability distributions. There is a quick reference guide that Simon Mak has made here: `https://github.com/resteorts/modern-bayes/blob/master/reading/distribution-quick-reference.pdf` that students have found helpful. Finally, I have prepared a summary of this information into one page, which is what you'll receive on exams. This can be found here: `https://github.com/resteorts/modern-bayes/blob/master/reading/common-distributions-one-pager.pdf`. I would suggest using this on homeworks or when practicing for the exams.
7. STA 360/601/602: I will assume that you are very familiar with this core class, so please brush up on this if you're feeling a bit rusty. You can find the most recent version of the course that I taught here `https://resteorts.github.io/teach/bayes20.html`.

Remark: This course will be tough without a strong foundation in the topics above, so please do review these in advance.

Expectations

1. Students are expected to learn github and how to use this before the first day of class as this is where all the course materials will be located.
2. Students are expected to be very familiar with R and are expected to know how to use R markdown.

3. All homeworks, reports, and take home exams (if applicable) should be submitted in Markdown .Rmd and .pdf format.

4. Please name your reports using the naming convention in the following example. As an example, please using the following naming convention **steorts-rebecca-homework1.Rmd**

5. All homework submissions must be made through **Sakai and Gradescope**.

6. **When submitting to Sakai, only one file must be uploaded.** Please zip together all materials for your homework assignment and upload the zipped file. As an example, please upload **steorts-rebecca-homework1.zip**, which should be a folder that contains **steorts-rebecca-homework1.Rmd** and **steorts-rebecca-homework1.pdf**.[1]

7. **When submitting to Gradescope, please only upload the rendered .pdf file.**

8. Your reports are expected to be reproducible and compile for full credit.

9. Students are expected to keep up with the reading in the course and have read before they come to class. Finally, if students find typos on the slides, please write them down with the slide and typo and give them to Professor Steorts for a timely correction to the course webpage.

10. **Attendance, either in real time or by watching recorded sessions promptly, is expected of all students given that homework and exam material will come from both class and lab.**

11. There will be between 3 – 5 homework assignments during the course of the semester.

Homework assignments will be announced on Sakai and Gradescope (along with the due date). Homework assignments should be **uploaded to Sakai and Gradescope (per the instructions above).** The homework assignments will be posted on the course github page.

All homework's involving analysis and code must be submitted to Sakai using Markdown and RStudio. Specifically, your homework must be reproducible. Your homework must be included as one file, therefore, please zip your files and submit all the files using a .zip extension. If you are unsure of how to do this, please see your TA or instructor during the first week of class during OH. Submissions via email to the TA's or instructor will not be accepted for credit.

Derivations for homework can be submitted in any format of your choosing as long as you convert this to a pdf file. Your work must be legible to the instructor and the TA's.

Recommendations that have worked well regarding derivations are Notability and Evernote. Other students have reported liking working with LaTeX within Rmarkdown.

If you have not downloaded LaTeX, you will need this in order for your .Rmd file to compile to a .pdf file. To install LaTeX, please see `https://www.latex-project.org/get/`. This will direct you to

[1] If you are working with data in a homework assignment, please make sure to also attach the data and also make sure that when you call the data in your markdown file, there are no hard coded commands. For example, make sure you do not set your working directory because we won't be able to reproduce your file.

options depending on if you are a Windows, Mac, or Linux user.[2] At least one student in the class has LaTeX working on Window using `https://tug.org/texlive/acquire-netinstall.html` or `https://miktex.org/download`. If you want to do a full install of LaTeX for Windows, see `https://tug.org/texlive/doc/texlive-en/texlive-en.html#installation`.

Prerequisites You are expected to have all pre-reqs to be in the course (see Prior Knowledge section above). Students are expected to be very familiar with `R` and are **encouraged** to have learned `LaTex` by the end of the course.

Required Textbook: There is no required textbook for this course.

Highly recommend supplementary text: *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Peter Christen, 2012, New York: Springer.

Optional supplementary text: *A First Course in Bayesian Statistical Methods*, Peter D. Hoff, 2009, New York: Springer. Please make sure you have a copy of the book as there will be required reading throughout the course. I will refer to this as "Hoff" throughout the course.

Optional supplementary text: *Statistical Inference, Second Edition*. Casella and Berger `https://fsalamri.files.wordpress.com/2015/02/casella_berger_statistical_inference1.pdf`

Optional supplementary text: *Some of Bayesian Statistics: The Essential Parts*. Rebecca C. Steorts, Copyright, 2015. `https://stat.duke.edu/~rcs46/books/bayes_manuscripts.pdf` I will refer to this as "PhD notes in the course."

Optional supplementary text: *Baby Bayes using R*. Rebecca C. Steorts, Copyright, 2016. `https://stat.duke.edu/~rcs46/books/babybayes-master.pdf` I will refer to this as "undergrad notes" in the course.

Optional supplementary text: *Bayesian Data Analysis*. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). CRC press.

The R Cookbook, `http://www.cookbook-r.com/`.

Github Setup and Commands: `https://www.youtube.com/watch?v=SWYqp7iY_Tc`.

Github Tutorial: `https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners`.

[2]For Mac users, I like this option here: `http://www.tug.org/mactex/`. Make sure to install the full version and not the small version. For a compiler, I like TexShop.

Grading Policy: The following grading policy will be used for this class.

Table 1: Grading Policy:

Homework	50%
Final Project	30%
Class Quizzes	20%

Individual homework extensions will only be given for extenuating circumstances. Given the current situation with the pandemic, please reach out early if you are not able to submit an assignment on time. Please contact Professor Steorts if you have an extenuating circumstance that prohibits you from completing the homework by the stated due date.

An overall score of s will result in a grade of:

A if $90 \leq s \leq 100$
B if $80 \leq s < 90$
C if $70 \leq s < 80$
D if $60 \leq s < 70$
F if $0 \leq s < 60$

or, for those taking the course on a Satisfactory/Unsatisfactory basis:

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

Exam dates cannot be changed and no make-up exams will be given. If extenuating circumstances prohibit you from taking an exam, please let Professor Steorts know before the start of the exam.

If you have an excused absence from an exam, the weight of the missed exam will be moved to the final exam. Please remember that you must take the final exam to pass the course. If extenuating circumstances prohibit you from taking the final exam, email Professor Steorts before the start of the exam.

Students are not to discuss any contents of any examination until after the exam grades are released back to them either in class or via Sakai. More specifically, students should not speak to anyone except the course instructor until exam grades are released to the entire class. This includes but is not limited to talking to other students, text, chat forums, and other means of communications where exam information could be shared to another student. Any student that does not follow this policy will be in violation of the Duke honor code.

Regrade Requests Regrade requests should be submitted through the regrade request form on Gradescope at `https://gradescope.com/auth/saml/duke`. Requests for a regrade must be made within a week of when the assignment is returned; requests submitted later will not be considered.

You should only submit a regrade request if there is an error in the grade calculation or a correct answer was mistakenly marked as incorrect. You should not submit a regrade to dispute the number of points deducted for an incorrect response. Please note that by submitting a regrade request, your entire assignment may be regraded and you may potentially lose points.

Due to the time consuming nature of responding to regrade requests, you should attend office hours and ask a member of the teaching team about the feedback before submitting the request. When you submit a request, please indicate which member of the teaching team you spoke with.

Note: Grades can only be changed by Professor Steorts. Teaching Assistants cannot change grades on returned assignments.

Late Work If there are extenuating circumstances that prevent you from completing an assignment by the **stated due date**, please let Professor Steorts know before the assignment is due. No late assignments will be accepted after solutions are posted or released to the rest of the class.

7 Important dates

1. Wednesday January 20: Classes Begin
2. Tuesday, February 2, Drop/Add Ends
3. Friday, March 5, Last Day for Reporting Midterm Grades
4. Tuesday - Wednesday, March 9 - 10 (No classes, short break)
5. Wednesday, March 24: Last Day to withdraw with a W (undergraduates only)
6. Monday, April 12: Wellness Day
7. Friday, April 23: Graduate and undergraduate classes end

A Additional Resources

There are many additional resources that are available and are included as an appendix.

R, github, and other course resources
Reviews and refreshers of both R and github can be found here: `https://github.com/resteorts/modern-bayes`. In addition, there are many other resources that are posted here that will help you with the course.

Academic Resource Center There are times students may need help with the class that is beyond what can be provided by the teaching team. In those instances, I encourage you to visit the **Academic Resource Center.** The Academic Resource Center (ARC) offers free services to all students during their undergraduate careers at Duke. Services include Learning Consultations, Peer Tutoring and Study Groups, ADHD/LD Coaching, Outreach Workshops, and more. Because learning is a process unique to every individual, they work with each student to discover and develop their own academic strategy for success at Duke. Contact the ARC to schedule an appointment. Undergraduates in any year, studying any discipline can benefit! Contact ARC@duke.edu, 919-684-5917.

CAPS Duke Counseling & Pyschological Services (CAPS) helps Duke Students enhance strengths and develop abilities to successfully live, grow and learn in their personal and academic lives. CAPS recognizes that we are living in unprecedented times and that the changes, challenges and stressors brought on by the COVID-19 pandemic have impacted everyone, often in ways that tax our well-being. CAPS offers many services to Duke undergraduate students, including brief individual and group counseling, couples counseling and more. CAPS staff also provides outreach to student groups, particularly programs supportive of at-risk populations, on a wide range of issues impacting them in various aspects of campus life. CAPS provides services to students via Telehealth. To initiate services, you can contact their front desk at 919-660-1000.

Technology Issues Students who may have limited access to computers and stable internet may request assistance in the form of loaner laptops and WIFI hotspots. For new Fall 2020 technology assistance requests, please go `https://duke.qualtrics.com/jfe/form/SV_bBdL1UL3N3iZNDD`. For returning students who wish to request an extension of a laptop or hotspot loan for Fall 2020 semester, please go `https://duke.qualtrics.com/jfe/form/SV_4SkFkWNrmvxhaBL`. For updates, please visit https://keeplearning.duke.edu/undergraduate-students/.

Accommodations for Remote Students If students cannot participate in synchronous or in-person course components (due to permanent time zone differences or temporary quarantine, for example), students should contact the instructor and academic dean to request an accommodation that will allow them to participate remotely. Please note that the experience may not be identical to that of local students.

Inclement Weather, Attendance, and Civic Engagement Policies Responsibility for class attendance rests with individual students. Since regular and punctual class attendance is expected, students must accept the consequences of failure to attend. However, in recognition of possible extra personal and academic stress this semester, I will grant excused absence, provided you discuss the absence with me and agree to make up missed work. Details regarding Trinity policies can be found here `https://trinity.duke.edu/undergraduate/academic-policies/class-attendance-and-missed-work`.

In the event of inclement weather or other connectivity-related events that prohibit class attendance, either in the location of the instructor or in the location of the student, I will notify you how we will make up missed course content and work. Asynchronous catch-up methods will likely apply.

If you have a situation regarding connectivity or any issue that prevents you from completing an assignment, please contact the instructor immediately. This will be dealt with on a case-by-base basis.

Technical and Zoom Support For technical help with Sakai or Zoom, contact the Duke OIT Service Desk at https://oit.duke.edu/help. You can also access the self-service help documentation for Zoom here and for Sakai here. The ARC (Academic Resource Center) has a student-friendly learning online guide and Zoom instructions here. Look on the sidebar on the left.

Mental Health and Wellness Resources If your mental health concerns and/or stressful events negatively affect your daily emotional state, academic performance, or ability to participate in your daily activities, many resources are available to you, including ones listed below. Duke encourages all students to access these resources, particularly as we navigate the transition and emotions associated with this time. Duke Student Government has worked with DukeReach and student advocates to create the Fall 2020 "Two-Click Support" Form, and Duke Reach has expanded its drop in hours as well.

1. DukeReach. Provides comprehensive outreach services to identify and support students in managing all aspects of wellbeing. If you have concerns about a student's behavior or health visit the website for resources and assistance. `http://studentaffairs.duke.edu/dukereach`
2. Counseling and Psychological Services (CAPS). CAPS services include individual, group, and couples counseling services, health coaching, psychiatric services, and workshops and discussions. (919) 660-1000
3. Blue Devils Care. A convenient and cost-effective way for Duke students to receive 24/7 mental health support through TalkNow. `bluedevilscare.duke.edu`

Managing daily stress and self-care are also important to well-being. Duke offers several resources for students to both seek assistance on coursework and improve overall wellness, some of which are listed below. Please visit `https://studentaffairs.duke.edu/duwell/holistic-wellness` to learn more about the following resources:

1. The Academic Resource Center: (919) 684-5917, theARC@duke.edu, or arc.duke.edu
2. DuWell: (919) 681-8421, duwell@studentaffairs.duke.edu, or `https://studentaffairs.duke.edu/duwell`
3. WellTrack: `https://app.welltrack.com/`