Almost All of Entity Resolution (STA 490/690)

Professor Rebecca Steorts

Welcome!

What is Entity Resolution?

Entity resolution (record linkage or de-duplication) is method used to remove duplicate entries from large, noisy databases that often do not have a unique identifer.

Can you think of some examples that you have seen?

Instructor

Prof. Rebecca Steorts

beka@stat.duke.edu
692 113 2482
Class: Friday, 10:00 AM -- 12:30 PM EDT
OH: Wednesday, 10:00 AM EDT -- 11:00 AM EDT (Tenative)

Teaching Assistants

Olivier Binette

olivier.binette@duke.edu
692 113 2482 (OH ID)
TBD

Where to find information

- Course website (all major course information here): https://resteorts.github.io/teach/er21.html
- Github (upload assignments): https://github.com/STA-690-S21
- Course syllabus https://github.com/resteorts/almost-all-of-er/blob/master/syllabus/syllabus-sta490-spring-2021.pdf

Remark: Videos on Github are too large to view, so you will need to download these if you wish to watch them. The fastest way, is to clone the repository and they will all download. They are stored this way in case your internet connection is slow.

Other resources

- Review of probability material: https://github.com/resteorts/modern-bayes/blob/master/reading/statistical-inference.pdf
- Simon Mak's Quick Guide to Prob. Distributions https://github.com/resteorts/modern-bayes/blob/master/reading/distribution-quick-reference.pdf
- A One Pager on Prob Distributions https://github.com/resteorts/modern-bayes/blob/master/reading/common-distributions-one-pager.pdf

Where can you find all the course information

Course website (all major course information here): https://resteorts.github.io/teach/er21.html

Prior Knowledge

- STA 210
  https://www2.stat.duke.edu/courses/Spring19/sta210.001/
- STA 230
  https://www2.stat.duke.edu/courses/Fall18/sta230/
- Linear algebra
  http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2(
- R programming (STA 199)
  https://www2.stat.duke.edu/courses/Spring18/Sta199/
- github (STA 199)
  https://www2.stat.duke.edu/courses/Spring18/Sta199/
- STA 360/601/602

Course Objectives

- Provide a foundation to entity resolution methods
- Explore, visualize, and analyze data in a reproducible and shareable manner
- Gain experience in data wrangling and munging exploring entity resolution models
- Work on problems and case studies inspired by and based on real-world questions and data
- Learn to effectively communicate results through written assignments and case studies

Unstable/slow internet?

Go to https://github.com/resteorts/almost-all-of-er/ and fork the repository.

- The videos are compressed, and this was done on purpose for those in the class that might have slow internet. I would suggest that everyone forks the repository to avoid any issues.
- Verify that you can play them (otherwise install something on your machine).
- Make sure to pull the repository each day. I update the repository very often as all the course resources are here (homeworks, lectures, data, videos). If you don't pull often, you might run into some issues!

Your Turn!

Create a GitHub account

Go to https://github.com/, and create an account (unless you already have one). After you create your account, click here and enter your GitHub username.

Tips for creating a username from Happy Git with R.

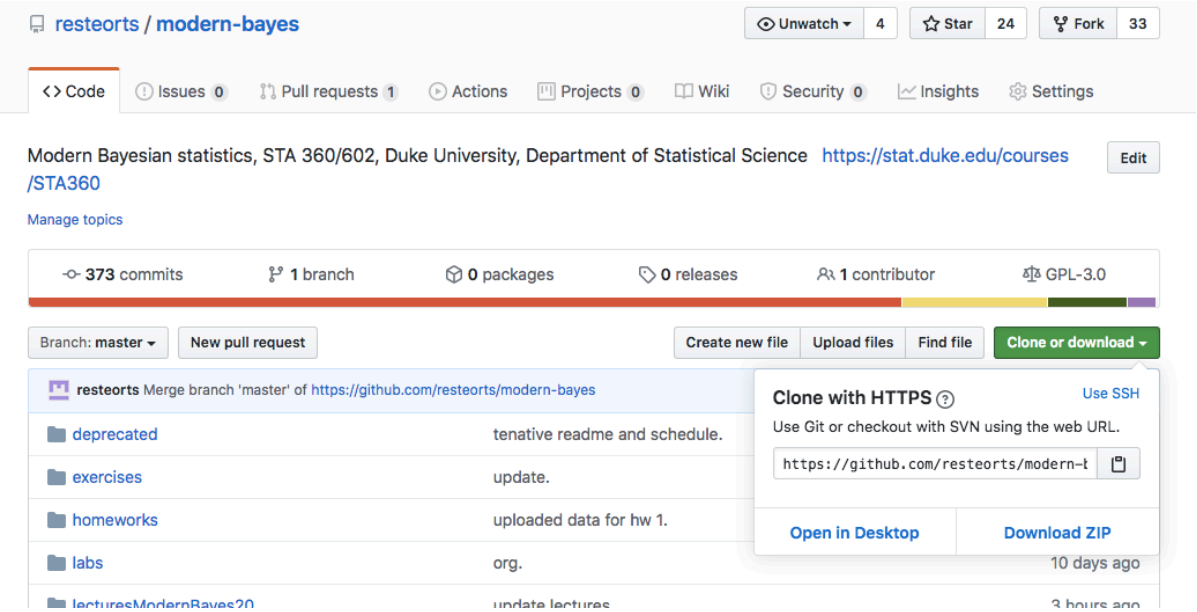
- Incorporate your actual name!
- Reuse your username from other contexts if you can, e.g., Twitter or Slack.
- Pick a username you will be comfortable revealing to your future boss.
- Shorter is better than longer.
- Be as unique as possible in as few characters as possible.
- Make it timeless. Don't highlight your current university, employer, or place of residence.
- Avoid words laden with special meaning in programming, like **NA**.

Login to github with your username (handle)

- Go to https://github.com/, and login with your github credentials.

Clone the class repository

- Clone the class repository **modern-bayes**



Raise your hand if you have any questions.

Clone the class repository

[Demo of cloning the class repository]

Discussion

Discuss the following with a partner (in a breakout room).

1. Start by introducing yourself! Name, year, major/ academic interest, favorite hobby.

2. Make sure that you can clone the repository and that your partner could as well.

3. Check to see that you're able to navigate the course webpage (in case that you lose access to the internet).

Tip: Make sure to pull from the repository quite often as the class webpage and materials will be frequently updated.

Course Policies

Class Meetings

Lecture

- Learn entity resolution methods
- Lectures will consist of learning methodology and applied coding techniques
- Lectures will be full of interactive exercises so make sure that you laptop is charged and your ready to do some computing!
- You might find it useful to have a tablet/document camera for taking notes
- An alternative to this is using your phone to take pictures using Evernote (students have said this worked well in the spring/fall). Please test things out in advance and make sure things are legible!

Supplemental Textbooks

- [Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Peter Christen]
    - Digital Copy or Hard Copy available for purchase.
- [A First Course in Bayesian Statistical Methods, Peter Hoff]
    - Free PDF available online through the library. Hard copy available for purchase.
    - Assigned readings on github.
- Some of Bayesian Statistics: The Essential Parts
    - Assigned readings.

Supplemental Textbooks

- Statistical Inference
- Baby Bayes using R
- Bayesian Data Analysis
- The R Cookbook
- Github Setup Video
- Github Tutorial

Activities & Assessments

- **Homework** : Individual assignments combining conceptual and computational skills. *Lowest score dropped.*
- **Quizzes** : Quizzes that will assess your knowledge of the course. *Lowest score dropped.*
- **Class**: **You will be responsible for keeping up with all class and lab material on a weekly basis (even though there are recordings)!**

Class Engagement

1. Students are expected to have read the assigned reading before coming to class.
2. Students are expected to have watched the pre-recorded videos (if available).
3. Prof. Steorts will go through the concepts in class again, highlighting the most important parts, providing clarifications, solutions/advice, and more advanced insights.
4. Prof. Steorts will provide exercises for the class to work through, an interactive environment, and discussions.
5. TAs will be present to help in these activities such that we can break into small groups and students can receive more individualized attention.
6. Please ask questions during class and being active as this will enhance the experience for everyone!

Homeworks

1. All code must be written to be reproducible in Markdown.
2. All derivations can be done in any format of your choosing (word, latex, markdown, written by hand) but must be converted to a pdf document. It must be legible.
3. All files must be zipped together and submitted to Sakai as one file (including Rmd and pdf). Your pdf document must be uploaded to Duke GradeScope. Please make sure to upload early to avoid issues.
4. Ask questions early if you have a problem to a TA regarding submission issues.
5. Your lowest homework grade will be dropped.

Remark: Sakai is for reproducible code. Gradescope is to make grading easier for everyone. Unfortunately, there is not a platform that handles both.

Please see the syllabus for all homework guidelines.

Why upload homework to Github Classroom

1. Github will be used for reproducibility checks
2. Github will be used to:
   - allow you to work in teams and collaborate using version control
   - return homework/exams more promptly
   - allow the teaching team to provide helpful feedback to yourself/your team

Quizzes

The format of quizzes will be announced in class and I will go over these will the entire class so that everyone can ask questions.

The goal of having quizzes is make sure that your baseline knowledge of entity resolution is solid before we move onto more advanced topics.

The goal of the quizzes is NOT to be stressful or overwhelming, but instead a positive, learning experience.

Grade Calculation

Component	Weight
Homework	40%
Quiz 1	13.33%
Quiz 2	13.33%
Quiz 3	13.33%
Final Project	20%

- See the syllabus for grade breakdowns.
- Grades will **never** be curved down.
- You are expected to attend lectures and labs in order to keep up with the course material.
- There will be no attendance grade or participation grade, but it's very important to attend class given that this is a small class!

Excused Absences

- Students who miss a class due to a scheduled varsity trip, religious holiday, or short-term illness should fill out the respective form.
    - These excused absences do not excuse you from assigned work.
- If you have a personal or family emergency or chronic health condition that affects your ability to participate in class, please contact your academic dean's office.
- Exam dates cannot be changed and no make-up exams will be given.

Late Work & Regrade Requests

- No late homeworks will be accepted, so please do not ask.
- No make up exams will be given.
- Regrade requests must be submitted within one week of when the assignment was returned.

Academic Honesty

All work for this class should be done in accordance with the Duke Community Standard.

> To uphold the Duke Community Standard:
>
> - I will not lie, cheat, or steal in my academic endeavors;
> - I will conduct myself honorably in all my endeavors; and
> - I will act if the Standard is compromised. Any violations will automatically result in a grade of 0 on the assignment and will be reported to Office of Student Conduct for further action.

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 or explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- On individual assignments, you may discuss the assignment with one another; however, 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.

Where to find 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.
- **Office Hours**: A lot of questions are most effectively answered in-person, so office hours are a valuable resource. Please use them!
- **Piazza**: Outside of class and office hours, any general questions about course content or assignments should be posted on Piazza since there are likely other students with the same questions.

Academic Resource Center

Sometimes you may need help with the class that is beyond what can be provided by the teaching team. In that instance, 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, 211 Academic Advising Center Building, East Campus – behind Marketplace.

Technology/Other

- Make sure that you have your zoom ids organized so that you're not late for class.
- Ensure the volume on all devices is set to mute.
- Refrain from engaging in activities not related to the class discussion. Browsing the web and social media, excessive messaging, playing games, etc. is not only a distraction for you but is also a distraction for everyone around you.
- If you have a question, I don't mind if you interupt me during class.
- If you find a **typo in the slides**, please write these down and email these to myself and Olivier so they can be fixed.

Accessibility

Please contact the Student Disability Access Office (SDAO) if there is an element of the course that is not accessible to you. There you can engage in a confidential conversation about the process for requesting reasonable accommodations.

Please note that accommodations are not provided retroactively, so please contact them as soon as possible. More information can be found online at access.duke.edu.

Inclusion

In this course, we will strive to create a learning environment that is welcoming to all students and that is in alignment with Duke's Commitment to Diversity and Inclusion. If there is any aspect of the class that is not welcoming or accessible to you, please let me know immediately.

Additionally, if you are experiencing something outside of class that is affecting your performance in the course, please feel free to talk with me and/or your academic dean.

Questions?

Announcements

Anything else here....

- **Please see me if you are on the waiting list**
- **If you're a student in the United States that cannot attend lectures/labs consistently due to being in a rural area or for any other reason, please see me after the first class.**
- **If your situation changes during the semester for any reason, please email myself and the TAs so that we can help you.**
- **In return, I would ask that everyone be flexible and understanding of everyone else (including instructors and TAs) as this is a very difficult and trying time.**