Lab #3: Record Linkage

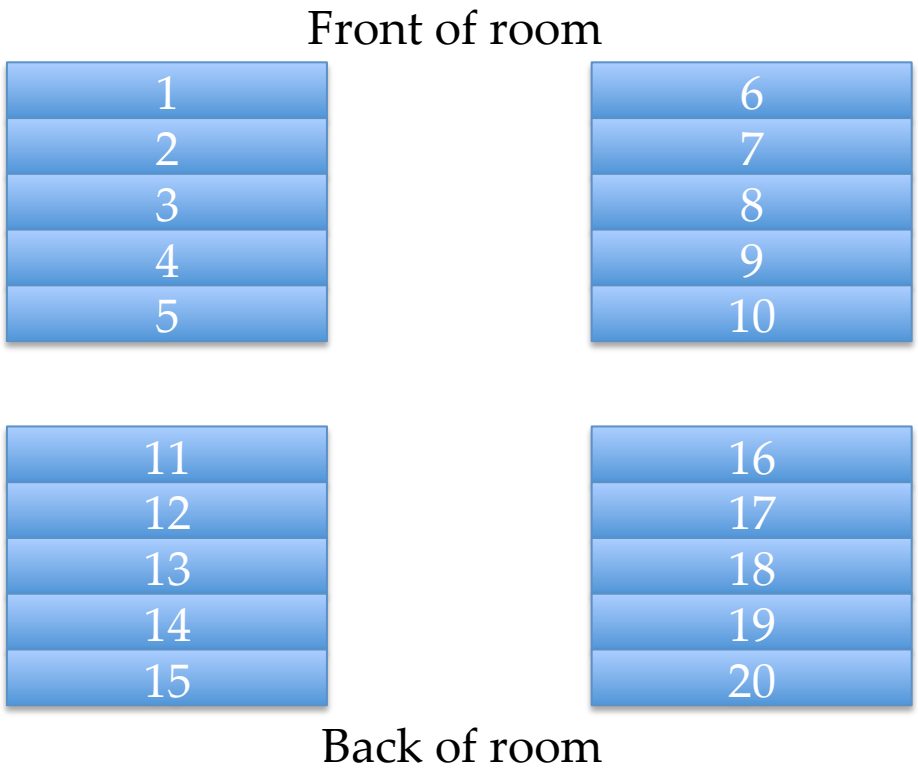
Everything Data

CompSci 216 Spring 2015

DUKE
COMPUTER SCIENCE

Seat assignment

See course website (under "Schedule") for team assignment



Format of this lab

- Discuss HW03 (amongst team members)
- More restaurant matching
- Team Challenge: Product matching

HW03 (5 min)

- Discuss Homework #3 with your teammates
  - Share your solutions and f1 scores

- At the end of this period, we will move onto the next part.

More Restaurant Matching (10 min)

- Now lets try your solutions on the **"full"** restaurant database!

- In addition to your solutions, we have selected 3 other sample solutions (from your submissions)!

More Restaurant Matching (10 min)

1. Did you get higher or lower f1 scores on the "full" restaurants database?

2. If the f1 score dropped, why did it drop?

3. Get your answers checked off with one of us before moving on to the next part.

Team Challenge: Product matching

- Match products from Amazon and Google

  amazon(id, title, description, manufacturer, price)

  google(id, title, description, manufacturer, price)

- How high an F1 score can you get?

- We will keep track of your scores on the board. You can make multiple submissions!

Part 1

	Small	Full
Match1.sql	0.9747	**0.9411**
Match2.sql	**1.0000**	0.9237
Match3.sql	0.9752	0.9130

Lesson Learned: Overfitting

- In most cases, you only have access to a sample of all records.
- Want your solutions to work even on unseen instances.
- More on overfitting in the next lab.

Lesson Learned: Text

- SQL is great for structured data
- Much harder to deal with unstructured fields like text.
- Need specialized methods for text processing
  - Two classes from now.

Finally

- Remember to **submit `team.txt` by midnight**
- Sample solutions to Lab #3 will be posted by tomorrow morning
  - Take some time on your own to go through the sample solutions