# CSE 4/587

## Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

## Day 27
## Final Review

# Announcements and Feedback

- Final Exam is on 12/14/22 @ 8AM in NSC 201
  - Tell me if you have an actual conflict ASAP
  - Exam **should** probably only take 2 hours, but you have the full 3 if needed
  - There will be assigned seating this time around
- Fill out course evaluations!
  - If 85% of evaluations are completed, everyone gets +1% on final grade

# Midterm Review

**Potential Topics:**

1. Classifiers
2. Naive Bayes
3. Logistic Regression
4. HIVE/Pig
5. Spark
6. Ethics in DIC
7. Misc/Previous Topics (sparingly)

# Classifiers [Lec 5, 6, 16-18]

1. Understand the classification of classifiers
2. Understand the development cycle of a classification problem
3. Understand the basics of the different classifiers we have discussed in class and how to use them
4. Understand the pros/cons of the classifiers discussed in class

# Naive Bayes [Lec 16-17]

1. Know the formulation of Bayes Law, and how to apply it to a given problem
2. Know how to take the application of many instances of Bayes Law and aggregate them into a single probability for the Naive Bayes model
3. Understand what Laplace Smoothing is, and what it addresses

# Logistic Regression [Lec 18]

1. Know what an odds ratio is
2. Know what the logit function is, and how to apply it to a given odds ratio
3. Know the final formula for logistic regression

# HIVE/Pig [Lec 19-20]

1. Have a basic understanding of how HIVE/Pig fit into the Hadoop ecosystem and what their purpose is within this ecosystem
2. Know the basics of how tables are divided and stored by HIVE/HBase
3. Understand high level concepts/components of HIVE/HBase such as Regions, RegionServers, META table, etc

# Spark [Lec 19-24]

1. Be able to read and understand Spark programs (in Python)
2. Understand what an RDD is, and how it is stored/computed in Spark
   a. Understand the difference between a transformation and an action
   b. Understand the difference between a narrow and wide dependency
   c. Know what a lineage graph is and what it is used for in Spark
   d. Be able to generate DAGs of RDD transformations
   e. Be able to divide DAGs of transformations into stages for execution
3. Understand the fault tolerance mechanisms used by spark
4. Understand the benefits Spark provides
5. Anything from the Spark Ungraded HW is also fair game

# Ethics in DIC [Lec 25]

1. Understand the different types of bias that may be part of our DIC applications
   a. Be able to explain what the types of bias are
   b. Be able to give examples of what may cause a particular type of bias to appear
   c. Be able to recognize situations that would cause a certain kind of bias to appear
   d. Be able to suggest possible solutions to address the different types of bias
   e. Understand which stages of the DIC pipeline each type of bias may appear in

# Misc/Previous Topics [Lec 26]

1. Have a basic understanding of topics covered by the midterm
2. Have a basic understanding of what was covered in Lecture 26 (course recap)