CSE 4/587
Data Intensive Computing

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

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Course Introduction

Course Staff

Professors

Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Shamsad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Teaching Assistants

Enjamamul Hoq
ehoq@buffalo.edu

Smarana Shrikant Pankanti
smaranas@buffalo.edu

Kartikeya Singh
ksingh35@buffalo.edu

To get to 208 Capen:
Take the elevators next to 1Capen to 2, then turn right.

Logistics

- **Course Website**
  - cse.buffalo.edu/~epmikida/teaching/sp23/cse487
  - All course materials, links, schedule, extra resources
- **Course Forum (Piazza)**
  - piazza.com/buffalo/spring2023/cse4587
  - All discussion for the course is hosted here – check regularly
- **UBLearns**
  - Assignment submission, grades

Logistics

- **Course Website**
  - cse.buffalo.edu/~epmikida/teaching/sp23/cse487
  - All course materials, links, schedule, extra resources
- **Course Forum (Piazza)**
  - piazza.com/buffalo/spring2023/cse4587
  - All discussion for the course is hosted here – check regularly
- **UBLearns**
  - Assignment submission, grades

Please keep class discussions on Piazza (private/anonymous posts exist)

Always include [CSE 487] or [CSE 587] in the subject line when emailing

Responsibilities

Attend lectures and participate

Read books and reference material

Attend office hours/Participate on Piazza

Complete the course project/assignments

Prepare for and take exams

Grading

Grade Breakdown:

- Homework: 10%
- Project: 40%
- Midterms: 20%
- Final Exam: 30%

Score (x)	Letter Grade	Quality Points
95% ≤ x ≤ 100%	A	4
90% ≤ x < 95%	A-	3.67
85% ≤ x < 90%	B+	3.33
80% ≤ x < 85%	B	3
75% ≤ x < 80%	B-	2.67
70% ≤ x < 75%	C+	2.33
65% ≤ x < 70%	C	2
60% ≤ x < 65%	C-	1.67
55% ≤ x < 60%	D	1
0% ≤ x < 55%	F	0

Academic Integrity

Collaboration, AI, Extra Resources

Do…

- Work together to brainstorm ideas
- Explain concepts to each other
- Discuss course content
- Include a list of your collaborators on all submitted work

Do Not…

- Write solutions when working together
- Describe the details of solutions to problems
- Leave your work in a place where it is accessible to another student

Collaboration, AI, Extra Resources

Do…

- Work together to brainstorm ideas
- Explain concepts to each other
- Discuss course content
- Include a list of your collaborators on all submitted work

Do Not…

- Write solutions when working together
- Describe the details of solutions to problems
- Leave your work in a place where it is accessible to another student

When in doubt, ask a member of the course staff!

Resource Policy

Do…

- Use materials provided by course staff (Piazza, Class, OH)
- Use materials from the course textbook or readings
- **Cite** all materials you reference for written work and code

Resource Policy

Do NOT…

- Reference random videos on YouTube that “helped you solve the problem”
- Hire “private tutors”
  - Save the money from Chegg
  - If you’re not doing the work yourself, you’re not learning
- Reference exact solutions found online

If you are caught using unauthorized resources, you get an F

Other Ways to Get an F

- Work in a group by assigning each person to a problem
- Copying your friend's homework because you forgot
  - Each homework is not worth a lot on its own
- Sharing your work with your friend
  - I have no way to know who did the work and who shared
- Submitting work without citations
  - Citing outside work will help you avoid AI repercussions
  - (we grade you on the work you did, but you won't get an AI violation)

Other Ways to Get an F

You are liable/punishable if someone else submits your work as their own.

Ways to Avoid an F

Don’t Cheat…

Ways to Avoid an F (amnesty policy)

Don’t Cheat…but we understand mistakes are made.

We will grant amnesty for any AI violation **IF** you tell us about it **BEFORE** we discover it

Why does Academic Integrity Matter?

Solutions may exist due to the simplicity of the problems

- Exercises try to force you to think a certain way
- Learning requires simplified/limited problems

You will not understand the design process from a solution

- Experience solving problems isn't obtained from reading solutions
- Anyone (who can read and write) can do copy-paste

Exact solutions to every problem don't always exist

- Stack Overflow/ StackExchange (and similar platforms) cannot do your job
- Open source solutions may not do what you need
- Depending on licensing, you can't always use open source solutions in closed source

Why does Academic Integrity Matter?

But it doesn't JUST hurt you...it also hurts the credibility of UB and its graduates!

How can you get the most from the course?

Be eager to learn about an emerging technology in high demand

Focus on opportunities to learn and grades will come naturally

Work hard to learn new skills and knowledge

Don’t be afraid to dive in and learn new languages/libraries

Be attentive in class

Work on the project yourself, even though teams are allowed

How should you assess success in this course?

Not by grade…

By new concepts you learn about data-intensive computing

By new skills you develop to solve data related problems

By new knowledge you gain about data applications, python libraries, MR, streaming data, etc.

…but do this and the grade will come too

What is the course about?

Foundational concepts in data intensive computing

Useful tools

Go from small data to big data

Go from big data to streaming data

▶

Identifying a problem

Data Acquisition

Understanding the data

Extracting features

Analysis

Visualizing

What is the course about?

Foundational concepts in data intensive computing

Useful tools

Go from small data to big data

Go from big data to streaming data

- Python
- Hadoop
- MapReduce
- Spark

What is the course about?

Foundational concepts in data intensive computing

Useful tools

Go from small data to big data

Go from big data to streaming data

Go from small, structured data, ie excel tables to big unstructured data (mainly text)

Big data data structures and algorithms (Hadoop, MapReduce)

What is the course about?

Foundational concepts in data intensive computing

Useful tools

Go from small data to big data

Go from big data to streaming data

New challenges with streaming data

What is it? Social media and enterprise data

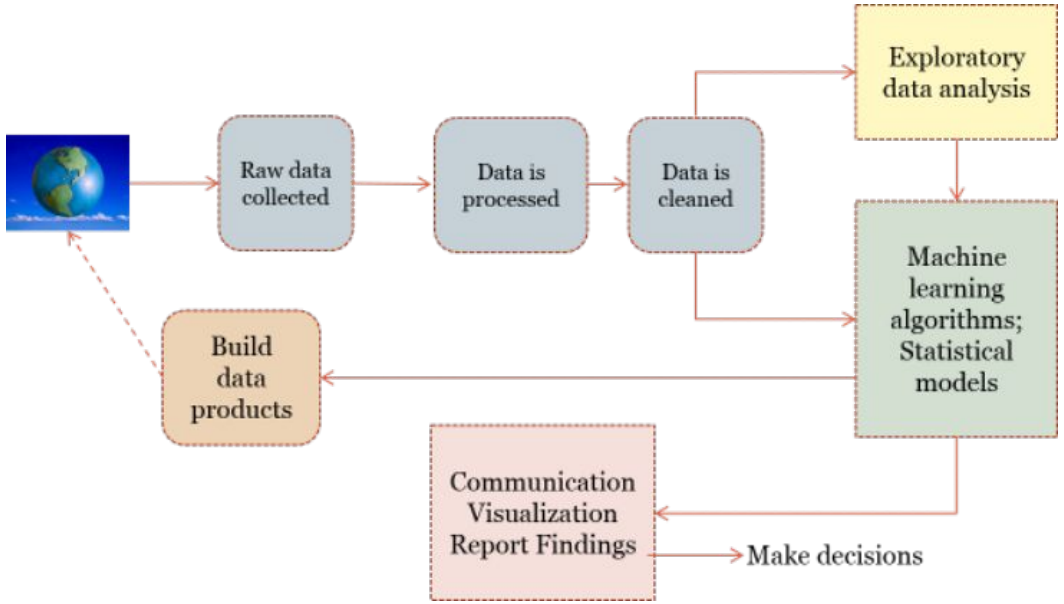
How do we characterize it and manage it (ie Spark streaming)

What will you learn?

Basic data analytics processes and how to apply them

Big data infrastructures and algorithms

Newer challenges (and how to handle them)



Questions?