CSE 4/587
Data Intensive Computing

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

Day 01
Course Introduction

Today’s Agenda

1. General course information/website/tools
2. Course content overview
3. Responsibilities as a student of the course
4. Assessing success in the course
5. First (small) homework

Course Information

Course Staff:

Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Enjamamul Hoq
ehoq@buffalo.edu

Smarana Shrikant Pankanti
smaranas@buffalo.edu

Course Website:

cse.buffalo.edu/~epmikida

There you'll find:

Syllabus
Piazza
Schedule
Slides
etc…

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

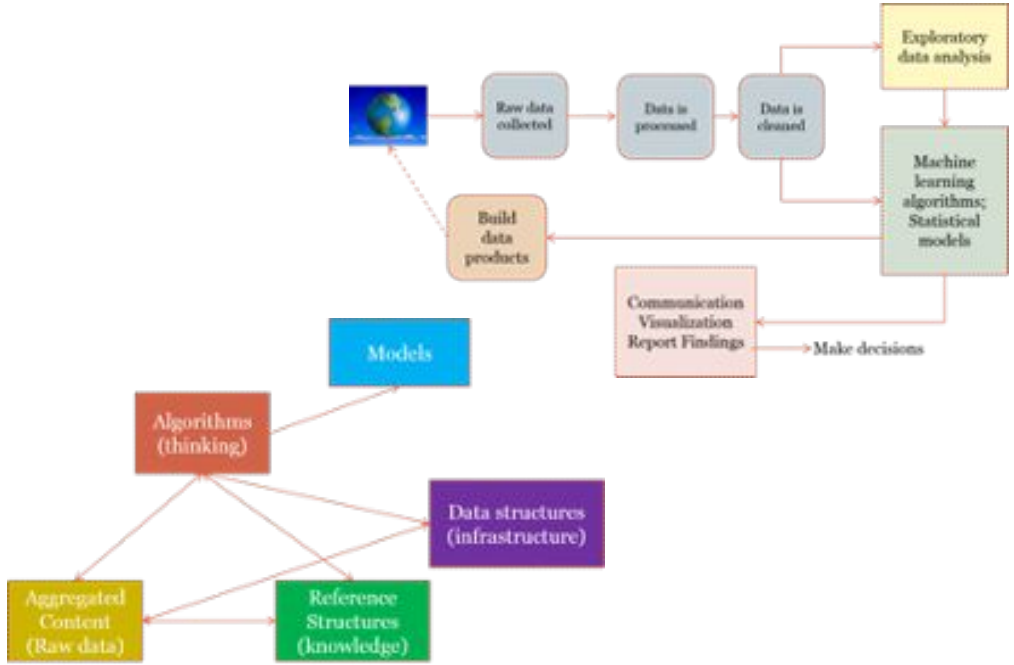
How do we characterize it and manage it (ie Spark streaming)

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). ie streaming data



What are your responsibilities as a student in the course?

Attend lectures

Participate

Read books and reference material

Attend office hours/Participate on Piazza

Complete the course project

Prepare for and take exams

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

TODO (by next week)

Read through chapter one in “Doing Data Science”

Form teams of 1 or 2 people

Start looking for a good data source for your project. Potential sources include:

- Pew, research (https://www.pewresearch.org/download-datasets)
- Data.gov
- Amazon and google datasets

Form a problem statement for your project along the lines of:
“I will analyze <this data set> to find out <something>, and address <this problem> with a data-driven solution”