What is Big Data?

- Things one can do at a large scale that cannot be done at a smaller one
- To extract new insights
- Create new forms of values

- Big Data is about analytics of huge quantities of data in order to infer probabilities
- Big Data is NOT about trying to "teach" a computer to "think" like humans
- Providing a quantitative dimension it never had before

The three (or four) Vs in Big Data

- **Volume**
  - Voluminous
  - It does not have to be certain number of petabytes or quantity.
- **Velocity**
  - How fast the data is coming in?
  - How fast you need to be able to analyze and utilize it
- **Variety**
  - Number of sources or incoming vectors
- **Veracity**
  - Can you trust the data itself, source of the data, or the process?
  - User entry errors, redundancy, corruption of the values
  - Data cleaning

Who is using Big Data?

http://www.cs.colostate.edu/~cs535  
Spring 2020 Colorado State University
Connected cars

- Single hybrid plug-in car generates up to 25 gigabytes per hour

- Connected cars
  - Traffic problem, re-routing based on the volume of traffic
  - Alerts driver when road conditions are hazardous by automatically activating anti-lock brake
  - Information shared by vehicles nearby

The Artemis project:
Saving “preemies” using Big Data

- The Artemis project
  - Dr. Carolyn McGregor

- Toronto’s Hospital for Sick Children, University of Ontario Institute of Technology and IBM

- Captures and processes patients’ data in real time
  - 16 different data streams
  - Heart rate, respiration rate, temperature, blood pressure and blood oxygen level
  - Around 3,200 data points per second
  - System detects subtle changes that may signal the onset of infection 24 hours before overt symptoms appear

http://www.cs.colostate.edu/~cs535
Look Who’s Peeking at Your Paycheck
- Experian’s Income Insight
- Estimates people’s income level
- Based on their credit history
- Trains the estimation model using selected credit history and tax information from IRS


Related research areas
- Storage systems
  - How can we efficiently resolve queries on massive amounts of input data?
  - The input dataset may be presented in the form of a distributed data stream
- Machine learning
  - How can we efficiently solve large-scale machine learning problems?
  - The input data may be massive, stored in a distributed cluster of machines
- Distributed computing
  - How can we efficiently solve large-scale optimization problems in distributed computing environments?
  - For example, how can we efficiently solve large-scale combinatorial problems, e.g. processing of large scale graphs?
Goal of this course

- Understanding fundamental concepts in Big Data Analytics
- Computing Systems + Scalable Algorithms and Models
- Learn about existing technologies and how to apply them

Computing systems

Algorithms and models

Analytics

Predictive models

Graph models

Storage systems and middleware

Computing frameworks

Specialized modeling tools

Communications

- Course Website
  - http://www.cs.colostate.edu/~cs535
- Announcements: Check the course website at least twice a week.
- Schedule (course materials, readings, assignments)
- Policies
- Canvas
  - Assignment submission
  - Grades
- Piazza
  - Discussion board

Contact Me

- sangmi@colostate.edu
- Office hour: Friday 10:00AM ~ 11:00AM and by appointment
- Office: CSB456
- URL: http://www.cs.colostate.edu/~sangmi

Contact GTAs

- Paahun Khandelwal
- Mohamed Chaabane
- Office hours (in CSB120 and online office hours TBA)

Research Group Meeting

When: 1:30-2:30pm Fridays
Where: CSB305

Course Structure

<table>
<thead>
<tr>
<th>Part A: Big Data Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Week 1 – Week 4</td>
</tr>
<tr>
<td>- Big Data Technology</td>
</tr>
<tr>
<td>- Purposes</td>
</tr>
<tr>
<td>- Hands-on experience</td>
</tr>
<tr>
<td>- Topics</td>
</tr>
<tr>
<td>- Introduction to Big Data</td>
</tr>
<tr>
<td>- Lambda Model</td>
</tr>
<tr>
<td>- Quick view of MapReduce</td>
</tr>
<tr>
<td>- Introduction to Apache Spark</td>
</tr>
<tr>
<td>- Analytics with Apache Storm</td>
</tr>
</tbody>
</table>

Course Structure | Part B: GEAR Sessions

- What is the GEAR Session?
  - Guided Exploration for Big Data Analytics Research
- Goals
  - Guided learning environment for advanced research topics in Big Data
  - Understanding different aspects of Big Data research with lectures and discussions
- Sessions
  - GEARS I: Peta-scale Storage Systems
  - GEARS II: Machine Learning for Big Data
  - GEARS III: Large Scale Recommendation Systems and Social Media
  - GEARS IV: Algorithmic Techniques for Big Data
- Duration: 2 weeks/session
  - Up to 3 lectures
  - 1 student-led research discussion

http://www.cs.colostate.edu/~cs535

Spring 2020 Colorado State University
### Course Structure | Part B: GEAR Sessions

- **Lectures**
  - Introducing fundamental concepts

- **Student-led discussions**
  - Students present research papers
  - 3 papers per session
  - We will have 5 sessions

- **GEAR is a team effort**
  - Each team will submit 1 critical review for each GEAR Session
  - Each team will lead 1 discussion session in this semester

### Course Component | Programming Assignments

- **Programming Assignment 1**
  - Implementing link analysis algorithm over the Wikipedia pages using Apache Spark
  - Due on 2/18 5:00PM
  - The description of PA1 is available now.

- **Programming Assignment 2**
  - Implementing real-time Twitter stream analysis using Apache Storm
  - Due on 3/10 5:00PM

### Course Component | Quizzes

- 7-8 online quizzes
- 48 hour window (with the 20 minute duration)
- Two lowest scores will be eliminated
- Quizzes will count 15% of total score of this course

### Course Component | Term Project

- **Objectives**
  - Students identify their topics for the term project
  - Students provide methodology to solve their problem
  - Students implement software solution
  - Students provide evaluation scheme for their software

- **Requirements**
  - Team effort
  - Deep learning aspect using TensorFlow
Course Component | Term Project

- Term project grading (40% of this course)
  - Term project planning: 1% (e.g. team members, a tentative title of your project)
  - Proposal document: 5%
  - Final paper: 24%
  - Final demonstration: 2%
  - Final presentation (peer review): 3%
  - Participation: 4%

- Late policy for the deliverable submissions
  - Up to a maximum of 2 day past the deadline.
  - 10% penalty per day will be applied

Course Component | Term Project

- Highlights of the Previous Term Projects
  - Supporting Emergency Response During Natural Disasters with Twitter Data
  - Winning Words in the Supreme Court
  - Rapid, Progressive Sub-Graph Explorations for Interactive Visual Analytics over Large-Scale Graph Datasets, (published and won the best paper award in IEEE/ACM International Conference on Big Data Computing, Application, and Technology 2019)
  - Mendel: A Distributed Storage System for Efficient Sequence Alignment and Similarity Searching (published in IEEE IPDPS 2016)
  - Processing Smart Grid Data In Real Time (DEBS grand challenge 2014)

Course Component | Team Efforts

- How can I be a GOOD team member?
- Evaluation categories
  - Frequency of participation in tasks
  - Task implementation
  - Communication
- Each category will be assessed by your peer teammates (including yourself)

Grading | Overview

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignments</td>
<td>20%</td>
</tr>
<tr>
<td>Assignment 1: 10%</td>
<td></td>
</tr>
<tr>
<td>Assignment 2: 10%</td>
<td></td>
</tr>
<tr>
<td>Term Project</td>
<td>40%</td>
</tr>
<tr>
<td>D0: Term project planning: 1%</td>
<td></td>
</tr>
<tr>
<td>D1: Proposal document: 5%</td>
<td></td>
</tr>
<tr>
<td>D2: Final paper: 24%</td>
<td></td>
</tr>
<tr>
<td>D2: Final demonstration: 2%</td>
<td></td>
</tr>
<tr>
<td>D3: Final presentation: 3%</td>
<td></td>
</tr>
<tr>
<td>Participation: 5%</td>
<td></td>
</tr>
<tr>
<td>Quizzes</td>
<td>15%</td>
</tr>
<tr>
<td>GEAR Sessions</td>
<td>25% (3% participation included)</td>
</tr>
</tbody>
</table>
# Grading | Overview

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignments</td>
<td>20%</td>
</tr>
<tr>
<td>Assignment 1:</td>
<td>10%</td>
</tr>
<tr>
<td>Assignment 2:</td>
<td>10%</td>
</tr>
<tr>
<td>Term Project</td>
<td>40%</td>
</tr>
</tbody>
</table>
| D0: Term project planning: 1%  
| D1: Proposal document: 5%  
| D2: Final paper: 24%  
| D3: Final demonstration:2%  
| D3: Final presentation: 3%  
| Participation: 5%       |
| Quizzes             | 15%                       |
| GEAR Sessions       | 25% (3% participation included) |

- 8% of these scores will be participation score
- 7% of these scores will be peer-evaluated team participation score

# Grading | Final Letter Grade

<table>
<thead>
<tr>
<th>Letter Grade</th>
<th>Total Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90.00 % and higher</td>
</tr>
<tr>
<td>B</td>
<td>80.00 ~ 89.99 %</td>
</tr>
<tr>
<td>C</td>
<td>70.00 ~ 79.99 %</td>
</tr>
<tr>
<td>D</td>
<td>60.00 ~ 69.99 %</td>
</tr>
<tr>
<td>F</td>
<td>Below 60.00 %</td>
</tr>
</tbody>
</table>

# Course Component | Course Policy

- No make-up for missed quizzes and exams
- Except for the case where student provided an advance written notice to the instructor based on an emergency
  - Supporting paper works will be requested
  - Two lowest quiz scores will be eliminated at the end of semester
- No Cell-phones in the class.
- No Laptops in the class.
  - If you need to use a laptop during lectures, please sit in the back row.
  - I will ask you to turn off your laptop if needed.
- Attend the class, ask questions, and discuss
- Check the course web page and canvas regularly
- Try new technologies and apply them
- Share your experiences with other students in class

Questions?

http://www.cs.colostate.edu/~cs535