Goals of this course

- Understanding fundamental concepts in Big Data
- Learn about effective Big Data technologies and how to apply them
  - Computing paradigms
  - Data retrieval methodologies
  - Data analytics patterns
  - Dataflow management for large applications
  - Scalable algorithms

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

What is Big Data?

The three 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
Data Engineers vs. Data Scientists, By Jesse Anderson
April 11, 2018, https://www.oreilly.com

• Estimate the population of London, which is then used to estimate the number of eligible ‘fighting men’ available to the king
• Dispel the myth that the plague was more rife in years of monarchy changes
• Investigate how London was expanding

Use all of the data, or just a little? [1/5]

Use all of the data, or just a little? [2/5]

• Estimating population
  - Using a 1658 map he assumed that there were 54 families living in each space of 100 yards square
  - Within the walls of London there were 220 such squares, making 12,180 families there
  - The number of deaths in London as a whole, both within and outside the wall, was four times the number of deaths within the wall alone

Use all of the data, or just a little? [3/5]

• John Graunt’s population example
  - “Usually [Graunt] explained his steps in solving problems, but he seldom included the actual calculations; and sometimes he omitted important information. Furthermore, his indirect approach sometimes went beyond the reliable use of his data, and the accuracy of some of his answers was difficult to evaluate.” — the Dictionary of Scientific Biography, Frank N. Egerton
  - How do we generate “The Life Tables” now?

Use all of the data, or just a little? [4/5]

• Conducting censuses
  - Costly and time-consuming
  - The US Constitution mandated one every decade
  - As the growing country measured itself in millions
  - By the late nineteenth century, even that was proving problematic
    - Data outstripped the Census Bureau’s ability to keep up.
  - The 1880 census took 8 years to complete
    - The information was obsolete even before it became available!
  - The 1890 census was estimated to require a full 13 years

Use all of the data, or just a little? [5/5]

• The Census Bureau contracted with Herman Hollerith
  - An American inventor used his idea of punch cards and tabulation machines for the 1890 census
  - Reduced duration to less than one year
  - Methods of acquiring and analyzing big data is very expensive
Now “Big Data” is available

- **Volume**
  - Every day 2.5 Quintillion bytes of data are created
  - 90% of world’s data today was created in the last two years

- **Velocity**
  - Every minute there are
    - 216,000 Instagrams
    - 204,000,000 emails sent
    - 277,000 tweets

- **Variety**
  - 90% of the data is unstructured
    - Tweets, photos, product information, etc.

In 2009 a new flu virus was discovered.

![Image of the newly identified H1N1 influenza virus](http://www.cdc.gov/h1n1flu/images.htm?s_cid=cs_001)

H1N1 virus

- Combines elements of the viruses that causes bird flu and swine flu
- Some projected outbreak to be on the scale of the 1918 Spanish flu that had:
  - Infected half a billion people
  - Killed tens of millions
- No vaccine was ready
- Public health authorities needed to know where it already was.

In the United States,

- The Centers for Disease Control and Prevention (CDC)
  - Requested doctors to inform the CDC about new flu cases
  - Tabulated the numbers once a week
  - Typically 1-2 week reporting lag
  - What are the problems with this system?

Detecting influenza epidemics using search engine query data (1/2)

  http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

Detecting influenza epidemics using search engine query data (2/2)

- Google
  - More than three billion search queries every day
  - Saves all the queries
- Looking at what people were searching for on the Internet
- Predict the spread of the winter flu in the United States
  - Nationally
  - Specific regions and states
How did they build a prediction model? (1/3)

- On the spread of seasonal flu between 2003 and 2008
- Aggregating historical logs of online web search queries
- Between 2003 and 2008
- Computed a time series of weekly counts for 50 million of the most common search queries in the US
- Weekly counts were kept for every query in each state

- Simple model that estimates the probability that a random physician visit in a particular region is related to an influenza-like illness (ILI)

How did they build a prediction model? (2/3)

- The probability that a random search query submitted from the same region is ILI-related
- \[ \text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \epsilon \]
  where:
  - \( I(t) \) is the percentage of ILI physician visits at time \( t \)
  - \( Q(t) \) is the ILI-related query fraction at time \( t \)
  - \( \alpha \) is the multiplicative coefficient
  - \( \epsilon \) is the error term
  - \( \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \)

How did they build a prediction model? (3/3)

- Each of the 50 million candidate queries was separately tested
- To identify the search queries that could most accurately model the CDC ILI visit percentage in each region
- Generated the highest scoring search queries

Evaluation of combining top-scoring queries

<table>
<thead>
<tr>
<th>Search query topic</th>
<th>Top 45 queries</th>
<th>Weighted n</th>
<th>Next 55 queries</th>
<th>Weighted n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza complication</td>
<td>11</td>
<td>18.15</td>
<td>5</td>
<td>3.60</td>
</tr>
<tr>
<td>Cold/flu remedy</td>
<td>8</td>
<td>5.05</td>
<td>4</td>
<td>3.08</td>
</tr>
<tr>
<td>General influenza symptoms</td>
<td>5</td>
<td>2.60</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>Terms for influenza</td>
<td>4</td>
<td>2.74</td>
<td>6</td>
<td>0.36</td>
</tr>
<tr>
<td>Specific influenza symptoms</td>
<td>4</td>
<td>2.54</td>
<td>6</td>
<td>3.74</td>
</tr>
<tr>
<td>Symptoms of an influenza complication</td>
<td>4</td>
<td>2.21</td>
<td>2</td>
<td>0.50</td>
</tr>
<tr>
<td>Antibiotic medication remedies</td>
<td>3</td>
<td>6.23</td>
<td>3</td>
<td>3.17</td>
</tr>
<tr>
<td>General influenza</td>
<td>2</td>
<td>0.16</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>Related disease</td>
<td>2</td>
<td>1.66</td>
<td>2</td>
<td>0.77</td>
</tr>
<tr>
<td>Antiviral medication</td>
<td>1</td>
<td>0.39</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td>Related disease</td>
<td>1</td>
<td>4.66</td>
<td>3</td>
<td>3.77</td>
</tr>
<tr>
<td>Unrelated to influenza</td>
<td>0</td>
<td>0.00</td>
<td>10</td>
<td>0.27</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>49.40</td>
<td>55</td>
<td>50.60</td>
</tr>
</tbody>
</table>

Most co-related queries

<table>
<thead>
<tr>
<th>Search query topic</th>
<th>Top 65 queries</th>
<th>Weighted n</th>
<th>Next 55 queries</th>
<th>Weighted n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated to influenza</td>
<td>0</td>
<td>0.00</td>
<td>10</td>
<td>0.27</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>49.40</td>
<td>55</td>
<td>50.60</td>
</tr>
</tbody>
</table>

What makes this work innovative?

- There have been other approaches:
  - CDC’s weekly report
  - Call volume to telephone triage advice lines
  - Tracking over-the-counter drug sales
  - Online activity for influenza surveillance
  - Search queries submitted to a Swedish medical website
  - Tracking search keys (“flu” or “influenza”) in Yahoo search queries

- What is the innovation in Google’s approach?
Part 0. Introduction
Course Introduction

Knowledge and Insights
Data, data, data… HOW?

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?

Related courses in CS at CSU
• Machine Learning/Statistics (CS480, CS545)
• Distributed Systems (CS455, CS555)
• Database Systems (CS430, CS530)
• Computer Communications (CS457, CS557)
• Computer Security (CS356, CS556)

Communications (1/2)
• Course Website
  • http://www.cs.colostate.edu/~cs435
  • Announcements: Check the course website at least twice a week.
  • Schedule (course materials, readings, assignments, recitations)
  • Policies
• Canvas
  • Assignment submission
  • Grades
• Piazza
  • Discussion

Communications (2/2)
• Contact Instructor
  • sangmi@colostate.edu
  • Office hours
    • Friday 10:00AM ~ 11:00AM and by appointment
  • Office: CSB456
  • URL: http://www.cs.colostate.edu/~sangmi
• Contact GTA
  • Bibek Shrestha
    • Office hours
      • Monday 6:00PM ~ 8:00PM (CSB120)
      • Friday 7:00PM ~ 9:00PM (CSB120)
Course Structure

1. Lectures
2. Assignments
3. Quizzes and Exams
4. Term Projects
5. Help Sessions

Assignments (1/2)
- Providing hands-on experience
- Assignment 0: Set up your Hadoop cluster and a simple example
  - Available on the course web
  - Due on AUGUST 31 5PM
- Assignment 1: Large scale data pre-processing using Hadoop MapReduce
- Assignment 2: Large scale data analytics using Hadoop MapReduce
- Assignment 3: Large scale data analytics using Apache Spark

Assignments (2/2)
- All assignments will be individual submissions
- Required programming language: Java
- Grading will be based on demo/interview
- Late submission policy
  - Up to 2 days of late submission is accepted
  - 10% deduction per day will be applied
- Assignment submission
  - Canvas

Course Structure

Lectures
- Week 1, 2
- Week 10
- Week 11, 12
- Week 13, 14
- Week 15

Assignments (1/2)
- Assignment 0: Set up your Hadoop cluster and a simple example
  - Available on the course web
  - Due on AUGUST 31 5PM
- Assignment 1: Large scale data pre-processing using Hadoop MapReduce
- Assignment 2: Large scale data analytics using Hadoop MapReduce
- Assignment 3: Large scale data analytics using Apache Spark

Assignments (2/2)
- All assignments will be individual submissions
- Required programming language: Java
- Grading will be based on demo/interview
- Late submission policy
  - Up to 2 days of late submission is accepted
  - 10% deduction per day will be applied
- Assignment submission
  - Canvas
Quizzes and Exams (1/2)

- Pop quizzes
  - About 10 quizzes this semester
  - Simple questions
  - Each quiz is about \(^{\sim}1\%\) of the course grade
  - 2 lowest scores will be eliminated at the end of the semester
  - Open book, open note
  - No Internet, no collaboration

Quizzes and Exams (2/2)

- Exams
  - Midterm and Final
  - 40% of the course grade
  - Mixture of simple questions and comprehensive questions

Course Structure

1. Lectures
2. Assignments
3. Quizzes and Exams
4. Term Projects
5. Help Sessions

Term Project

- Comprehensive Big Data Analytics experiences
- Phase 0: Find your teammate! (August 30, 2018 By 5:00PM)
- Phase 1: Term Project Proposal
- Phase 2: Final submission: Software and Report
- Phase 3: Presentation (Class presentation) and demonstration of your software
- Term project is a team effort 3-4 students per team
- No single member team allowed!
- http://www.cs.colostate.edu/~cs435/

How do I select my Term Project topic?

- Do you have a topic in your mind?
  - Yes
  - No
- Did you find interesting topic?
  - Yes
  - No
- Did you come up with your own topic?
  - Yes
  - No
- Do you have teammate?
  - Yes
  - No

Example topics from previous courses

- Similarity and Clustering on The Million Song Dataset
- Methods Used to Analyze Eclipse and Mozilla Bug Data
- Movie Recommendations using Collaborative Filtering
- Climate Visualization and Predictive Analysis
- Wikipedia Page Traffic Statistics Analysis
  - Trending Topics and Page Count Prediction on Wikipedia Traffic Log Data
- Supporting Emergency Response During Natural Disasters with Twitter Data
Example topics from previous courses

- "Time to Answer" for Questions on stackoverflow.com using Map Reduce
- Analysis of words for spell-checking in search queries using digitized books and articles
- Trend Analysis To Improve Donations
- Trends in baby names and predicting popular names
- Regression Analysis of New York City Taxi Fares
- Predicting Stock Similarity Using K-Means clustering
- Who is Building Wikipedia?

Course Structure

1. Lectures
2. Assignments
3. Quizzes and Exams
4. Term Projects
5. Help Sessions

Recitation (Optional)

- Friday 3:00pm ~ 3:50PM, CSB130
- Recitation will be led by GTA and recorded
- Hands-on sessions

Grading Policy

<table>
<thead>
<tr>
<th>Category</th>
<th>Points</th>
<th>Percentage of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total: 30</td>
<td></td>
<td>30%</td>
</tr>
<tr>
<td>PA0: 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1: 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2: 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3: 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term Project</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total: 20</td>
<td></td>
<td>20%</td>
</tr>
<tr>
<td>Phase 0: 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 1: 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 2: 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 3: 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quizzes and participation</td>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>Midterm Exams (20% x 2)</td>
<td>Total: 40</td>
<td>40%</td>
</tr>
</tbody>
</table>

Final Letter Grade

Letter grades will be based on the following standard breakpoints

- >= 90 is an A,
- >= 89 is an A-
- >=88 is a B+
- >=80 is a B,
- >=79 is a B-
- >=78 is a C+
- >=70 is a C,
- >=60 is a D,
- and <60 is an F

Course Timeline
Course Policy

• No make-up for missed exams
  • Except in extraordinary circumstances (e.g., major illness, family emergency)

• No make-up for missed quizzes
  • Except for the case where there is an emergency

Plagiarism

• We use plagiarism detection software

• Please do not copy code or sentences from your references or your colleague’s work

• Discuss with each other, but do not write your software together!

More importantly,

• Attend the class, ask questions, and discuss
• Check the course web page, Piazza, and Canvas regularly
• Try new technologies and apply them
• Share your experiences with other students in class

Questions?