CS 730R: Topics in Data and Information Management

Big Data Analytics
Today

• Introduction
• Topics overview
• Course logistics
Instructor/Coordinator

- Li Xiong, http://www.mathcs.emory.edu/~lxiong
- Research areas: data privacy and security, data integration and analytics
- Office hours: TuTh 4-5:30pm or by appointment, WSC E412
Meet everyone in class

• Introducing yourself
  – Name, year
  – Research interests
What the class is about

• Seminar course
• Read papers on emerging topics in data and information management
• Theme: big data analytics
Why Big Data

• We live in big data
  – Digital data, sensor data, surveillance data, scientific data ...
A day in the internet

- 294 billion emails are sent
- 2 million blog posts
- 172 million different people visit Facebook
  - 532 million statuses are being updated
  - 250 million photos are uploaded
- Twitter: 40 million
- LinkedIn: 22 million
- 22 million hours of TV shows and movies are watched on Netflix
- 864,000 hours of video are uploaded to YouTube
- 18.7 million hours of music is streamed on Pandora
- 35 million apps are downloaded
- 2 million search queries per minute on Google
- ....
Why Big Data

• New opportunities
  – Improve decision making from business, medicine, to homeland security
  – e.g. IBM “Watson”; 1000 genomes project
CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.

IT COMES FROM EVERYWHERE. IT KNOWS ALL.

ACCORDING TO THE BOOK OF WIKIPEDIA, IT'S NAME IS "BIG DATA."

BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.

BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.

LET US PAY.

IS IT TOO LATE TO SIDE WITH EVIL? SHHHH! IT HEARS YOU.
Why Big Data

• National "Big Data Initiative" by the White House, March 2012
  – six Federal departments and agencies committing more than $200 million to Big Data research projects
  – Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data
Why Big Data

• Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it ...

– Dan Ariely, Alfred P. Sloan Professor of Behavioral Economics, MIT, Jan 6, 2013
– 385 shares, 1342 people like this
What is Big Data - Wikipedia

- Big data is a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
- Large
  - Moving target: terabytes to petabytes
- Complex
- Challenges: capture, curation, storage, search, sharing, analysis
What is Big Data – MIKE 2.0

• The **degree of complexity** within the data set
• The amount of value that can be derived from innovative vs. non-innovative analysis techniques
• The use of longitudinal information supplements the analysis
What is Big Data – MIKE 2.0

• Number of independent data sources and permutations (aircraft sensors)
• Difficult to delete individual records – privacy a common concern (toll road data)
• Big data - big complexity rather than big volume
• Big data can be very small (aircraft sensing)
• Not all large datasets are big (media streaming)
What is Big Data – Gartner

• 3V model
  – Volume (amount of data)
  – Velocity (speed of data)
  – Variety (range of data types and source)

• Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization
Challenges

• Huge data volume
• Integration across heterogeneous data sources
• Real-time analytics, longitudinal analytics, streaming data
• Uncertain data, probabilistic data
• Data security and privacy
Computing paradigms

• Cloud computing
• Crowdsourcing, human computation
Focus/Scope

- Data integration
- Real-time analytics, longitudinal analytics, streaming data
- Uncertain data, probabilistic data
- Crowdsourcing
- Data security and privacy

- Not focused on: text data, social media and social networks
- Not focused on Fundamentals
Papers

• Database, data mining conferences and journals
• General computing journals (e.g. CACM)
• Not: distributed computing, Information retrieval, NLP, ...
Goals

• Learn about what people are doing on big data
• Identify new problems and solutions
• Relate to classic problems and solutions in database, data mining, and data privacy and security
• What is new ... what is old
Non-technical goals

• Learn/practice how to write annotated bibliography/paper critiques
• Learn how to present a paper and lead discussions
• Write a research survey (term paper) – hopefully serve as literature review for part of your dissertation research
• Identify new research topics/interests
Workload/Grading

• Paper reviews (1 per week) 50%
• Paper presentations (~3) 20%
• Class discussions 10%
• Term paper 20%
Paper review

• 1 page
• NOT just a summary of the paper, but your critical opinion of the paper
• Point out limitations, extensions, or interesting applications of the ideas
• Connect and contrast the paper to what we have read
Sample Format

1. Summary
   - What is this paper about (50,000 feet)?
   - How does it build upon, extend, or differ from other work?
   - What are the key ideas in the paper; What is novel or interesting IN YOUR OPINION?

2. Positive/Strong Points [What I like about this paper. Provide at least three bullet items.]
   - S1
   - S2
   - S3
   Example: "system Q handles the common case well -- very popular file access requests are given greater resources using a novel credit scheme"

3. Negative/Weak Points [What I don't like about this paper. Provide at least three bullet items.]
   - W1
   - W2
   - W3
   Example: "the computing model assumes that we know in advance when users will be logging in to their machines, this is not realistic!"

4. Research Questions and Points for Discussion [be ready to ask and discuss them in class. Examples include clarification of specific issues (at most one item of this type please), questions about claims made by the authors, `what if` questions about situations that aren't handled well by the ideas in the paper, or specific suggestions for improvement.]
   - D1
   - D2
   - D3
   Example: "however, if we had an approximate model of how long a machine might be available, we could recast the problem in this way: ...."
Presentations

• Each class will have 1-2 paper presentations, depending on the depth/content of the paper
• You will have ~3 presentations in total
• I will take note of longer presentations, particularly if the paper is more complex, and you make effort to understand the paper details by reading other background papers.
Presentation Format

• Context and background
• Key technical ideas and results
• It’s relation to other papers or ideas
• Positive/Negative points (and why)
• Discussion points and research questions
Presentations

• Keep it interesting
  – don’t want gory paper details nor total fluff

• Research/discussion questions
  – Go beyond the claims of in the paper
  – Limitations, extensions, improvements

• You may find/use the ppt online but
  – Put it in your own words
  – Understand everything you are presenting
Term paper

• Comprehensive paper reviews on a sub-topic
• Place the work in context
• Categorize, compare and contrast the work
• Taxonomy
• Point out open questions and future directions
Late Policy

• Paper reviews/term paper
  – 2 late review allowances, each can be used to turn in a late review or term paper within a week

• Paper presentations
  – Please give a week’s notice if you need to reschedule
Schedule

• Reading list posted online; roughly one subtopic every 2-3 weeks
• Paper review due every Tuesday
• Paper presentations assigned on Wednesday morning for the following Tuesday and Thursday
• Reading list for term paper due mid-semester
• Term paper due end of semester
related courses

• CS554: Database Systems
  – database design
  – complex queries, data storage and indexing
  – XML, spatial, and temporal data management
  – ontologies and semantics
  – parallel data management and MapReduce

• CS 584: Topics in Computer Science: HPC & Parallel Processing