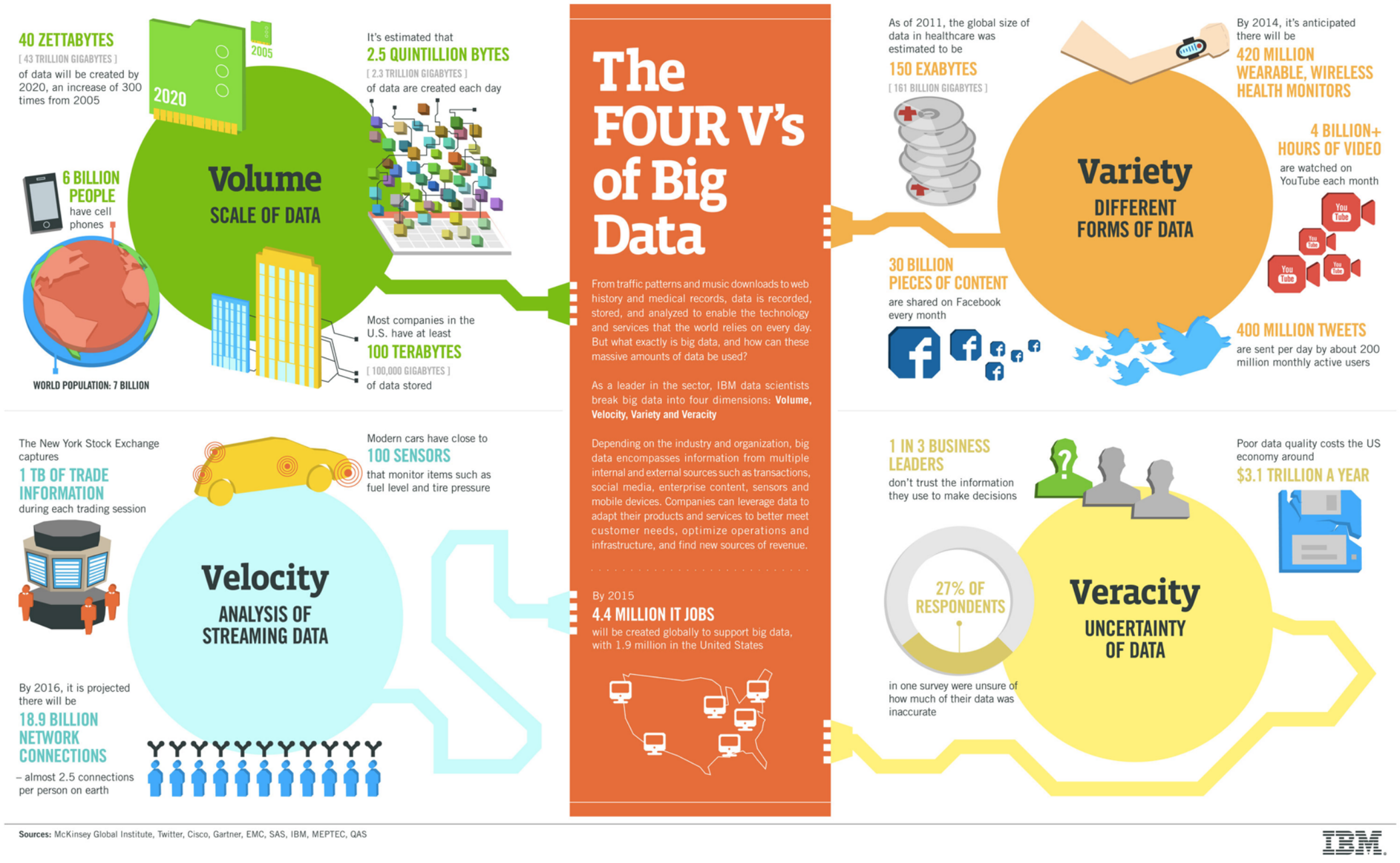
Introduction and Course

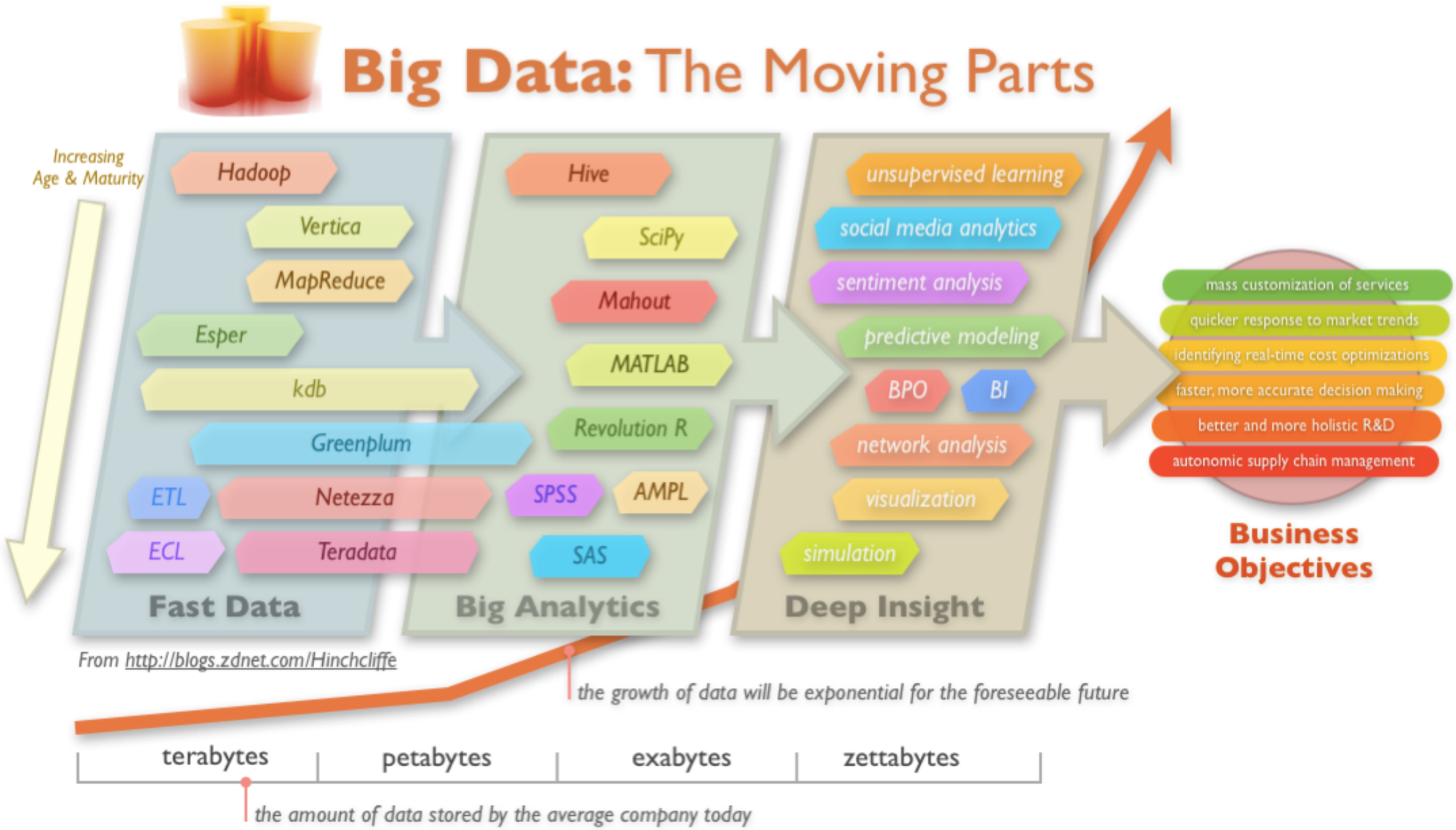
CS 584: Big Data Analytics

4 V's of Big Data



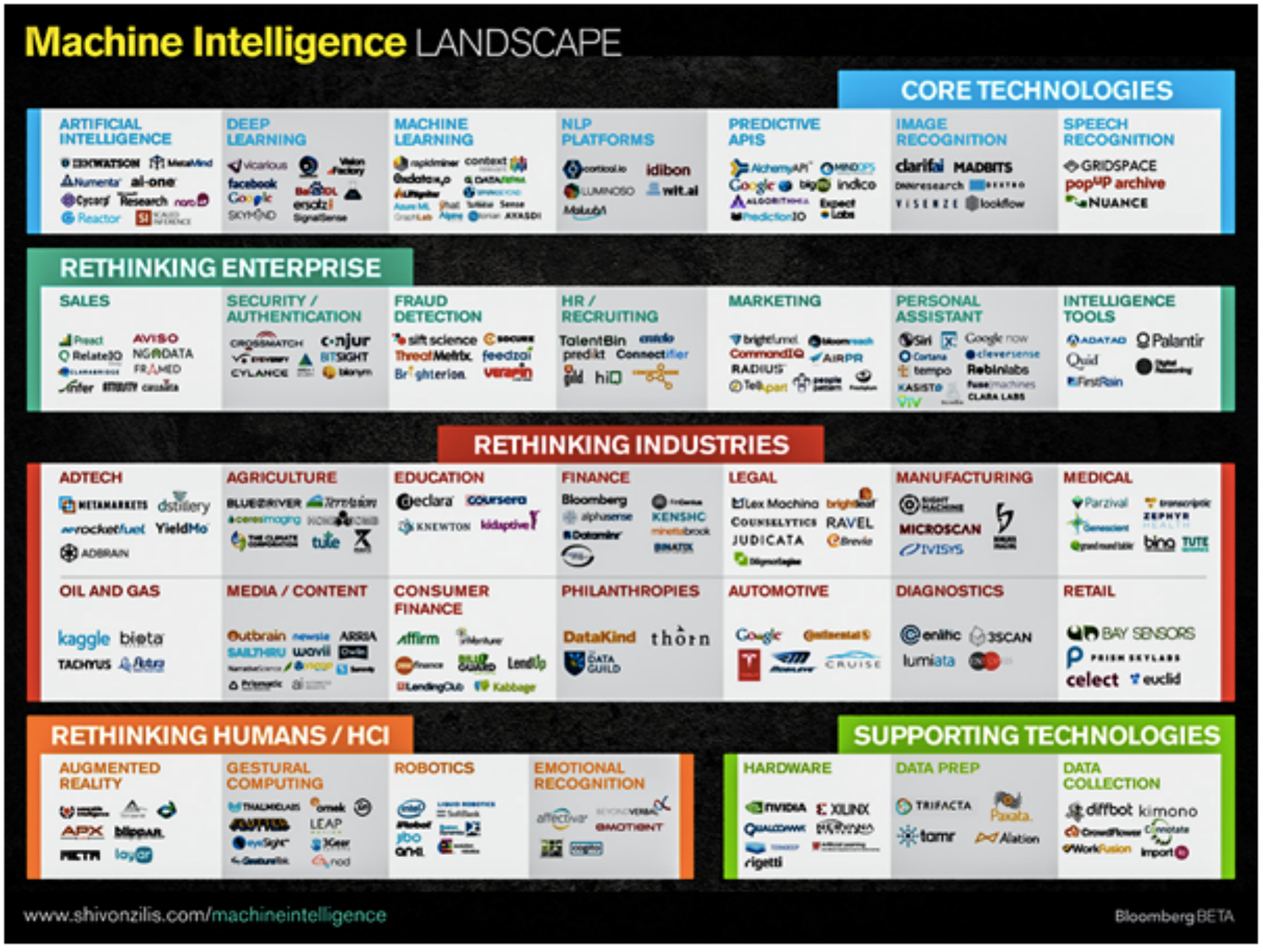
http://www.ibmbigdatahub.com/infographic/four-vs-big-data

What is Big Data Analytics?



http://www.zdnet.com/i/story/60/39/001648/big_data_the_moving_parts_large.png

Current State of Machine Learning



http://www.shivonzilis.com/machineintelligence

Course Objectives

- Learn about the various techniques to analyze big data
- Present and lead class discussion for at least one of the papers listed in the schedule
- Identify strengths and weaknesses in existing research via written critiques (reviewer practice)
- Develop your portfolio of projects for internships and jobs (can result in a potential paper)

Course Overview

- Scalable machine learning and data mining algorithms
  - Large-scale optimization techniques
  - Random projections and hashing
  - Streaming and sketching algorithms
  - Distributed matrix factorization
  - Tensor factorization
- Webpage: http://joyceho.github.io/cs584-s16/index.html

About Me

- Undergraduate / MEng from MIT
- PhD from University of Texas at Austin
- Research interests:
  - Data Mining / Machine Learning
  - Healthcare Informatics
- Email: joyce.c.ho@emory.edu
- Office Hours: Tues/Thurs 1-4 pm @ MSC W414 or by appointment
- More information: http://joyceho.github.io

Course Format

- Structured more as a seminar course
- First few weeks of class, traditional lecture format with slides posted the night before online
- Afterwards, will move towards class presentations (with some lectures sprinkled in between) where 1 student leads
- Cover approximately one paper per class

Class Presentations

- Each student should at least skim the paper before each class for better discussions
- One student leads the class discussion
- Submission of slides at least two days before your presentation is scheduled
- Meet with me to discuss your plan ahead of time so we can iterate at least one
- Can use slides that may already exist for the paper

Course Project

- Work in groups of 2-3
- Emphasis on public data sets (e.g., Kaggle competitions, MovieLens, KDD Cup, etc.)
- Open-ended: almost anything will work as long as it relates to data mining and machine learning
- Project proposal due by spring break for feedback

Grading

Class Presentation	25%
Paper Reviews	15%
Course Project	45%
Participation	15%