Introduction to Data Science

CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

Course Objectives

- Introduce
  - the theoretical foundations, algorithms, and methods of deriving valuable insights from data
- Study
  - big data management and processing techniques, data analytics, statistical methods and models, data visualization, and etc

Goal of the Class

- Define and explain the key concepts and models relevant to data science.
- Design, implement, and evaluate the core algorithms underlying an end-to-end data science workflow.
  - E.g., the experimental design, data collection, mining, analysis, and presentation of information derived from large datasets.
- Apply "best practices" in data science, including facility with modern tools (e.g., Hadoop).

Mapped objectives in IDEA

- Learning fundamental principles, generalizations, or theories
- Learning to apply course material (to improve thinking, problem solving, and decisions)
- Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course

Class Topics

- Data Exploration
- Data Preprocessing
- Mining and Analytics
- Visualization
- Evaluation
- MapReduce
- Cloud Computing
- …

Course Structure and Administrivia

Course Information

- Instructor
  - Kyumin Lee
  - kyumin.lee@usu.edu
  - Office: MAIN 401D
  - Office hours: 10:15-11:15am, T/R or by appointment
- TA
  - Sravan Rudrayagari
  - rsravan94@gmail.com
  - Office: MAIN 422
  - Office hours: 10:15-11:15am, M/W or by appointment
- Class hours:
  - 9:00am ~ 10:15am TR
  - MAIN 406

Course Information

- Course web page
  - http://digital.cs.usu.edu/~kyumin/cs5665/
  - Check frequently
- Sign up our Google Groups
  - https://groups.google.com/d/forum/cs5665-fall2016
- Our group maliing list
  - cs5665-fall2016@googlegroups.com

Course Materials

- No primary textbooks required
- References
  - Data Mining
  - MapReduce
  - Visualization

Course Communication

- The website (especially, schedule page) will be updated often
  - Check it regularly
- I will email important announcements and post them to the website
- You may email me anytime ... but I only guarantee a response within four days
- The best way to discuss general questions or share something cool stuff is to email it to our google group.

Class Structure

- Lectures
  - By instructor -- I'll teach fundamental data structures and algorithms
  - By us - Discussion and interaction in the class
- Your part
  - Homework
    - 4 assignments
  - Practicum
    - Each week we will tackle some practical application of Data Science -- be it a tool, a framework, or some other artifact that will help you transition your theoretical foundation into practice.
  - Midterm
  - Project
    - Proposal, execution, workshop presentation
- Participation
  - Ask good questions

Grading

- 5% Attendance and In-class discussion
- 32% (four) Assignments
- 20% Midterm
- 13% Practicum
- 30% Project

Assignments

Assignments

- 4 assignments
- Submit your solution to Canvas
  - You only use Canvas for submitting your assignments
- Late day policy: look at the syllabus

Midterm

- The midterm exam is closed book.
- You may bring one standard 8.5" by 11" piece of paper with any notes you think appropriate or significant (front and back).
- No electronic devices allowed.

Practicum

- You will choose a tool or framework to introduce
  - I will provide a list of tools or frameworks soon
- Your job is to install it and, run and test it with a sample dataset.
- Tell us what issues you had, sticking points, and other insights that can help us.
- Prepare and present slides (MAX 10 mins)

Project

The Project

- 2 or 3-person team
- Project idea:
  - Propose anything you wish
  - You are encouraged to talk to me
- **30% of your final grade!!**

Project Grading Criteria

- [25%] Project Proposal:
- [25%] Check Point
- [50%] Project Workshop: Dec 6 and 8 in-class

Homework 0 [0% of your final grade]

- Register for the Google Group
- Then go to the group and post a message in each of four threads (I've already started the threads for you):
  - An introductory message. Who are you?
  - What you think "data science" is.
  - What you expect to get out of this class. Be specific.
  - A link to a cool example of data science in action with a brief (one sentence) explanation of why you picked the link.
- Due: Sept 4, 2016 by 11:59pm

To download lecture notes

- ID: cs5665
- Password: science

What’s Next

- Form a team by Sept 13 and notify team members to me
- Do Homework 0
- Read the Readings

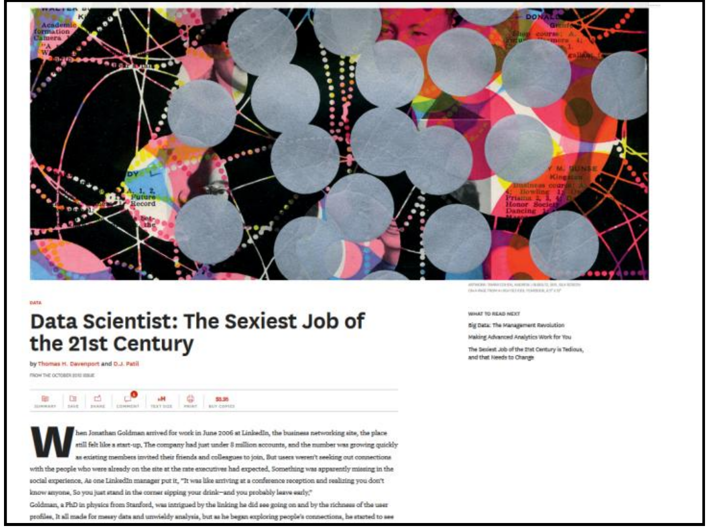
Data Science…

Data?

How Big?

Big Data in 2011

- 1,200,000,000,000,000,000,000 bytes of data generated in 2011
- Facebook - 1,150 million users
- Gmail – 425 million users
- Skype – 300 million users
- Twitter – 500 million users (200M active)
- WhatsApp – 300+ million users
- Youtube – 1,000 million users (4 B daily views)
- Instagram - 150 million users
- Waze – 50 million users
- Amazon – 209 million users
- Ebay - 120 million users
- Paypal - 132 million users
- Google searches – ~12 billion (monthly, US alone)



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see

WHAT TO READ NEXT

Big Data: The Management Revolution

Making Advanced Analytics Work for You

The Sexiest Job of the 21st Century is Tedious, and that Needs to Change

What is Data Science?

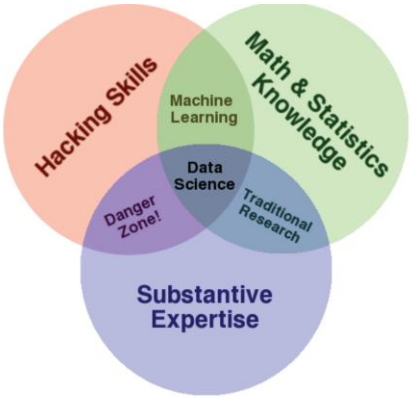
http://www.quora.com/What-is-data-science

- Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.
- But data science is not merely hacking, because when hackers finish debugging their Bash one-liners and Pig scripts, few care about non-Euclidean distance metrics.
- And data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a ^A delimited file into R if their job depended on it.
- Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible.

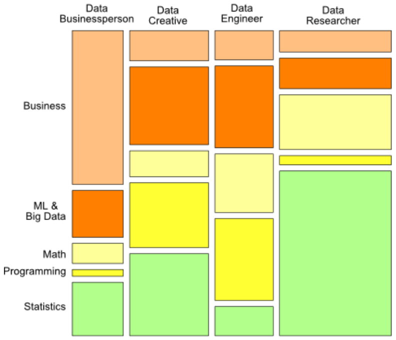
Nathan Yau
“Sexy Skills of Data Geeks”

- Statistics - traditional analysis you're used to thinking about
- Data Munging - parsing, scraping, and formatting data
- Visualization - graphs, tools, etc.

http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/



http://drewconway.com/zia/?p=2378



- Surveyed from over 250 data scientists

http://www.datacommunitydc.org/blog/2012/08/data-scientists-survey-results-teaser/

Data Science…

- Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of data analysis fields such as data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD)

https://en.wikipedia.org/wiki/Data_science