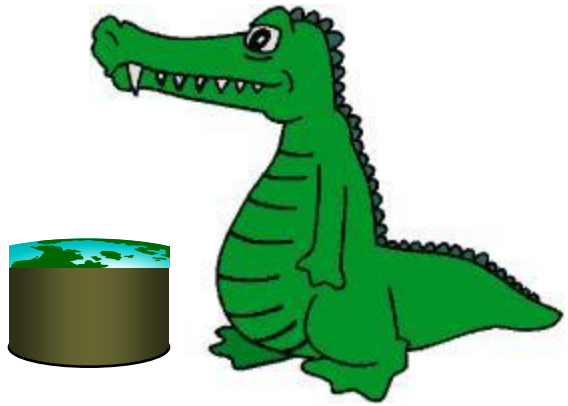
CAP4770/5771
Introduction to Data Science
Fall 2015

University of Florida, CISE Department
Dr. Daisy Zhe Wang



(Big) Data Science

Data science is the study of the generalizable extraction of knowledge from data. It incorporates varying components and builds on techniques and theories from many fields with the goal of extracting meaning from data and creating data products. [Wikipedia]

By 2018, the U.S. faces a shortage of 1.5 million analysts with Big Data know-how. [McKinsey]

- Data -> Knowledge -> Action
- UF Data Science Research: *Probabilistic Knowledge Base construction (DBMS + SML)*

Course Goals

- Teach state-of-the-art tools to do Data Science
  - Data Collection, Storage, Manipulation, Querying and Processing
  - Data Analytics and Modeling
  - Big Data and Parallel Processing
  - Tell the Story through Data
  - Data Science Applications

Vital Information

- Instructor: Prof. Daisy Zhe Wang
- Office: E456
- Class time: Mon/Wed/Fri 3-3:50pm (8th period)
- Office hours: Wed 3:50-4:50pm/Fri 2-3pm or by appointment
- TA: Xiaofeng Zhou, Miguel Rodriguez (Office hour: TBA)
- Course page will be up later this week:

https://ufl.instructure.com/courses/320501

(read announcements frequently!)

Course Formats

- Lectures
  - 28 Lectures on 14 topics
  - Guest Lectures (up to 5)
- Labs and Homework
  - 1 bootcamp on unix/python (a.k.a., lab0)
  - 4 Monday Labs
- In-class Midterm and pop Quizzes
  - Review lectures
- Final Project
  - System and algorithm development
  - Presentations & write-ups

This Course will Teach the following through Lectures and Hands-on Labs

- Perform data collection and preparation.
- Program for data analytics in Python.
- Apply statistical modeling, machine learning for structured and unstructured data analysis.
- Process data at scale using map-reduce over cloud services.
- Use visualizations, presentation and write-ups to tell your data story.

This Course will NOT

- Teach the basic programming (e.g., JAVA) and data structures
- Teach some of the advanced topics in Data Science (e.g., probabilistic graphical models)
- Attempt to improve existing Data Science systems and algorithms

Other Data Science courses @ UF

- First of the three-course series in the Data Science curriculum, followed by
  - Projects in Data Science (CAP4773/CAP6779)
    - Consider extending your final project to a semester-long project
  - Advanced Topics in Data Science (CAP 6769)
    - Research Oriented: paper reading, presentation, research projects

Pre-requisites

- Require
  - Data Structures and Algorithms (COP3530)
  - Or equivalent
- Prefer
  - Information and Database Systems I (CIS4301)
  - Statistics and Probabilities (STA 5325/5328)
  - Programming experience with JAVA, SQL, R, Python
- Academic honesty

Course Outline

- An introduction to the basic data science techniques including programming in Python, SQL/SPARQL and Map-Reduce for small and big data manipulation and analytics.
- Teach basic techniques for data collection, data preparation, data querying, data analytics including pattern mining, classification, clustering, data visualization, and parallel computing platforms.
- Teach advanced data analytics techniques including NLP, knowledge extraction, graph analytics, graph querying, knowledge bases and crowd sourcing.
- Introduce key application areas of data science including business intelligence, social media, biomedicine, and e-discovery.
- More details: https://ufl.instructure.com/courses/320501

Suggested Readings

- Mining of massive datasets, A. Rajaraman and J.D. Ullman, Cambridge University Press, 2011. ISBN-10: 1107015359, ISBN-13: 978-1107015357, http://www.mmds.org/ (public online access)
- Doing Data Science, Cathy O'Neil and Rachel Schutt, O'Reilly Media Inc., http://proquest.safaribooksonline.com/9781449363871
- Python for Data Analysis, Wes McKinney, O'Reilly Media Inc., http://proquest.safaribooksonline.com/9781449323592
- Data Science and Big Data Analytics, EMC/WILEY, ISBN: 978-1-118-87613-8

Textbooks and Software Required

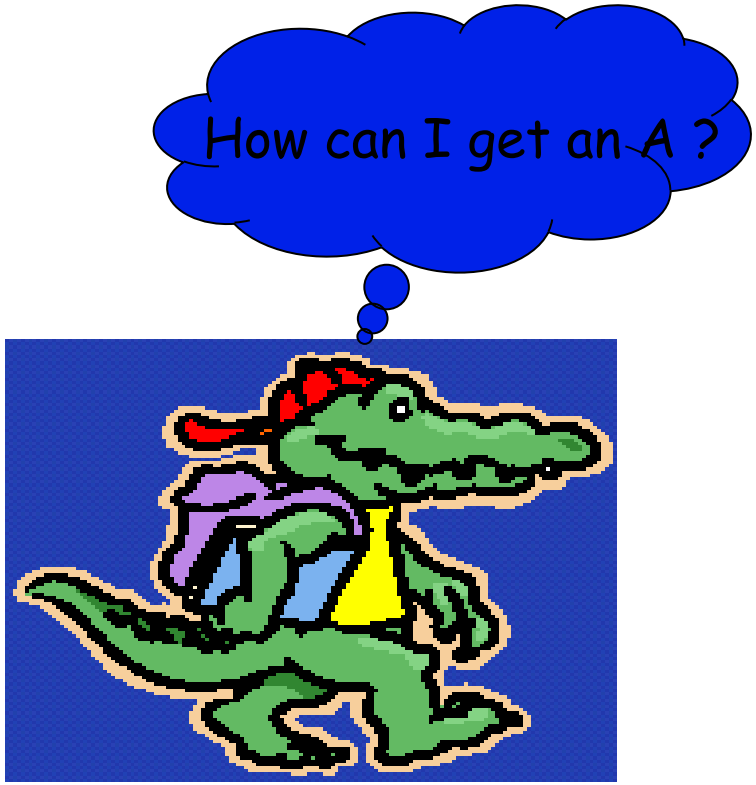
- None – refer to recommended reading online and class materials including lecture notes, labs, homework, quizzes

- We will be using Amazon Web Services (AWS) and software supported on top of AWS. AWS credits will be given to each student.

Attendance and Expectations

- **We require class attendance and participation, since most of the class material will be delivered in class.**
- **Moreover, in-class pop quizzes will be conducted to test the understanding of the material via canvas.**
- **Personal laptops/tablets/smart phones are required in class for participation and pop quizzes.**
- Please return your labs/homework/projects in time. Late returns will cause 20% deduction in your grade for that lab/homework/project for each late day.

Course Evaluation



- 2 In-class Midterms (30%)
- 5 In-class Labs & Homework (35%)
- Final Project (25 %)
- In-class Pop Quizzes (10%)

- Late submission for project and homework: 20% per day.

Computing Resources

- Amazon Web Services
  - need a credit card to create an AWS account
  - 100$ AWS credits per student will be provided
  - Will be used for Lab/Homework 3 & Final Project
  - Should be enough to complete the projects
  - Beyond the credit limit is at your own cost

Lab 0-3 (25%)

- Lab 0 (Sep 2&4): Unix & Python bootcamp – 7%
- Lab 1 (Sep 14): Panda – 7%
- Lab 2 (Sep 21): NLTK – 7%
- Lab 3 (Oct 5): map-reduce – 7%

Lab 3

- Work in groups of ~2 people
- Get AWS started
- Finish AWS tutorials on AWS account and S3 setup, create and run a job flow, command line tools, AWS instance types and pricing, EMR, debugging, etc.
- Implementation of a well-defined algorithm over a given dataset using Map-Reduce on AWS

- Evaluation: correctness, performance and selected code review

Lab 4 and onward

- Lab 4 (Oct 26): Scikit – 7%
- 2 midterms (Oct 14, Nov 13), 2 review lectures (Oct 12, Nov 9) – 30%
- Final project (25%)

Final Project (25%)

- Work in groups of ~6 people
- Given datasets with guidelines for analytics

- Evaluation: demo, presentation and write-up on data processing, analytics techniques applied and data product results

In-class pop quizzes (10%)

- In the form of timed question answering via canvas
- Personal labtop/tablet/smart phone needed to take pop up quizzes
- Grade made up of attendance and correctness of answer

Grading

Roughly the boundaries will be:

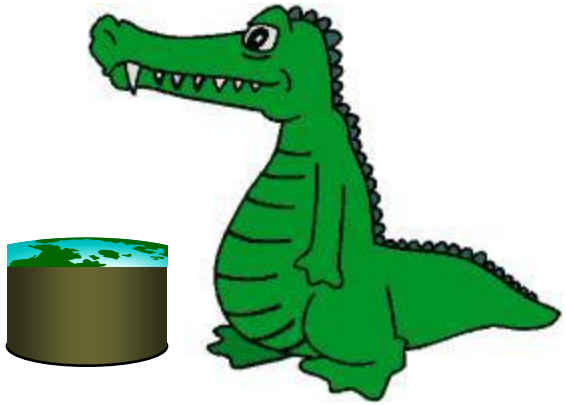
- 90 -- 100 A
- 86 -- 89 B+
- 80 -- 85 B
- 76 -- 79 C+
- 70 -- 75 C
- 66 -- 69 D+
- 60 -- 65 D
- 0 -- 59 E

The boundary for A-, B-, C- will be decided at the end of the semester.

Questions?

Next Lecture – Overview of Data Science

If time permits: A taste of Data Science Project



NIST Data Science 2015 Fall Pre-Pilot Evaluation Plan

- National institute of standard and technology (NIST) http://www.nist.gov/
  - The institute's official mission is to:

    Promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.