COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

Discussion Section A01

Teaching Assistant (TA): Abdullah

Instructional Assistant (IA): Kyra

Hi, I’m Abdullah

- 1st Year MS Data Science student in HDSI
- **Email**: aashfaq@ucsd.edu
- **OH**: Thursdays 1-2pm (zoom)
- Academic Background:
  - Undergrad Electrical Engineering (2020) with focus on hardware for ML
- Work Background:
  - Data Scientist at Afiniti (Telecom + Banking Industry)
  - Data Scientist at Totogi (Telecom Industry)
  - Incoming Data Engineer Intern (Summer 24) at Tesla
- Current Interests:
  - Discovering novel uses of ChatGPT
  - Tennis* (Just started)

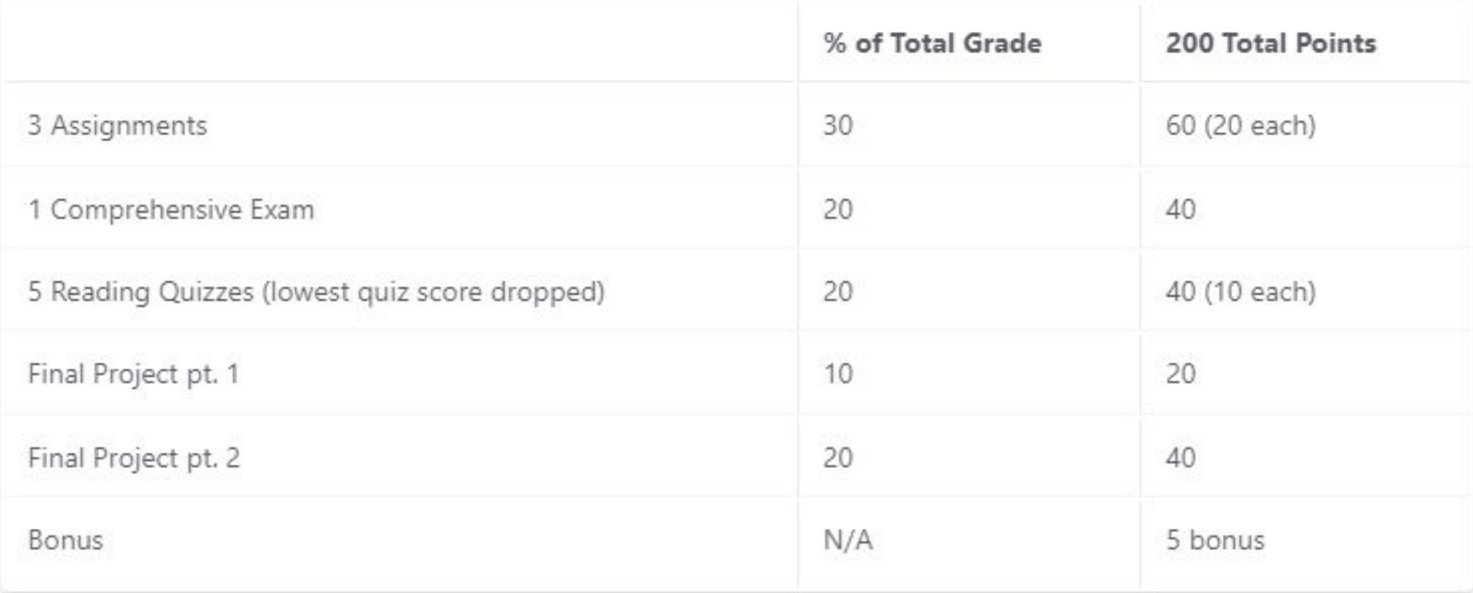
Hey! I’m Kyra

- 2nd year Data Science Student minor in ASL
- **Email**: kkdeng@ucsd.edu
- **OH**: Mondays 4-5pm (zoom)
- Currently on board for CASA and in VSA

Logistics

- Master link: https://kshannon-ucsd.github.io/cogs9/ (Everything you need is available)
- Discussion section attendance is optional
- Sections are not recorded; however, the slides will be uploaded to Drive/GitHub and the link to the same will be provided in the master link
- Quizzes, exams and assignments are to be done individually
- Final project is to be done as a group (Split into 2 parts)
- Please try and reach out to us on piazza or email
- Make good utilization of office hours
- DO NOT CONTACT THE PROFESSOR DIRECTLY UNLESS IT IS PERSONAL/EMERGENCY

Grading



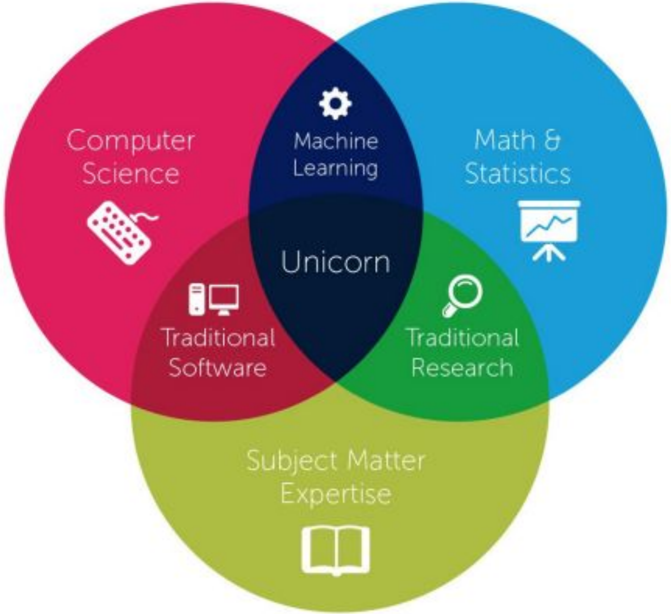
	% of Total Grade	200 Total Points
3 Assignments	30	60 (20 each)
1 Comprehensive Exam	20	40
5 Reading Quizzes (lowest quiz score dropped)	20	40 (10 each)
Final Project pt. 1	10	20
Final Project pt. 2	20	40
Bonus	N/A	5 bonus

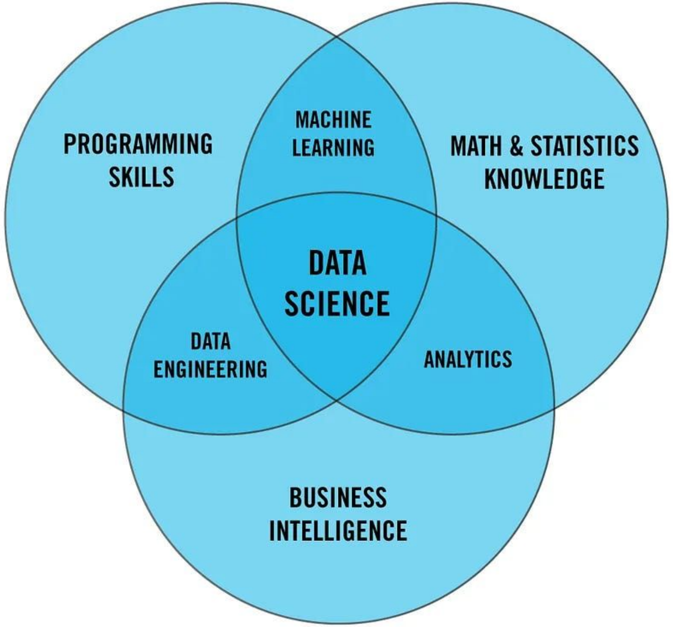
Discussion Sections Outline: Mostly Hands-on

- Week 2: Introductions, Making teams, Reading 1 (Part 1)
- Week 3: Reading 1 (Part 2), Python Basics with Jupyter Notebook
- Week 4: Reading 2, Getting data and wrangling it using Pandas
- Week 5: Reading 3, Assignment 1, Basics of programming for data science
- Week 6: Reading 4, Final Project Part 1 reviews/discussions
- Week 7: Assignment 2, Data Visualization and EDA demo
- Week 8: Assignment 3, Machine Learning demo
- Week 9: Reading 5, Closing thoughts
- Week 10: Final Project Part 2 reviews/discussions

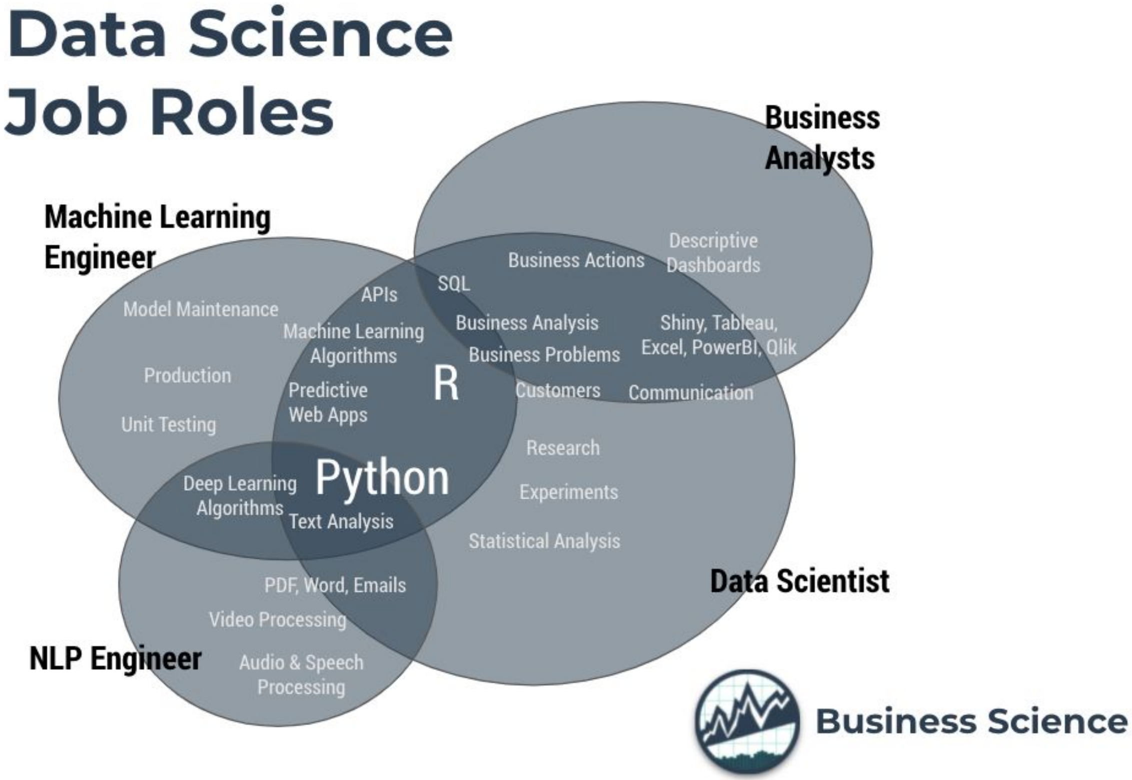
Careers in Data

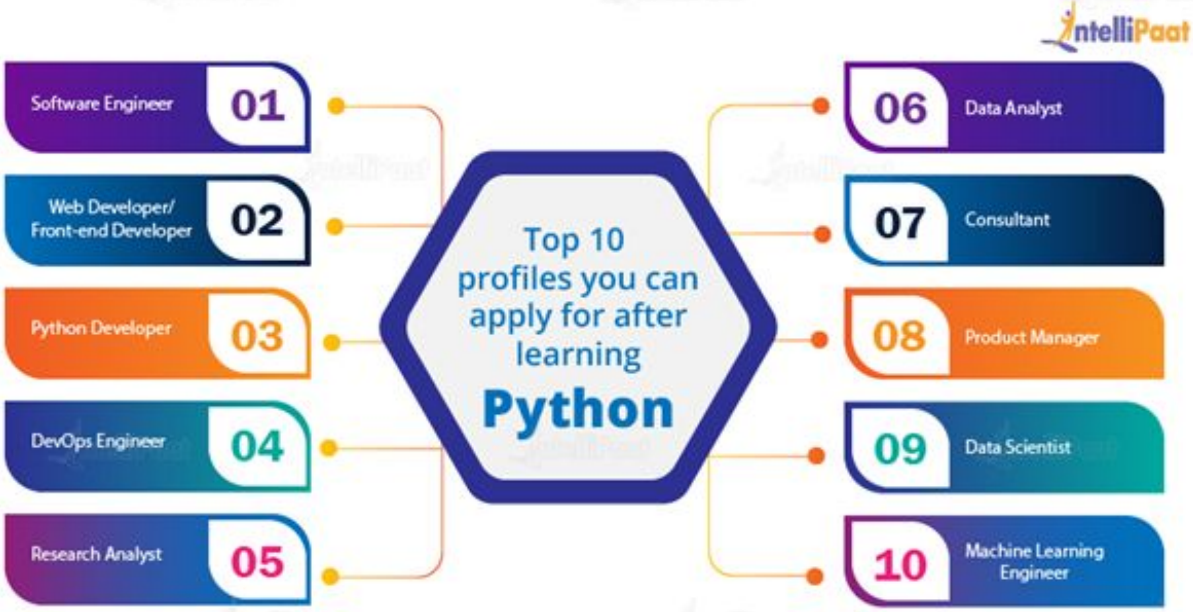
Careers in Data





Skills


Data Science
Job Roles
Business
Analysts
Machine Learning
Engineer
Model Maintenance
APIs
SQL
Business Actions
Descriptive
Dashboards
Machine Learning
Algorithms
Business Analysis
Shiny, Tableau,
Excel, PowerBI, Qlik
Business Problems
Production
R
Predictive
Web Apps
Customers
Communication
Unit Testing
Research
Deep Learning
Algorithms
Python
Experiments
Text Analysis
Statistical Analysis
Data Scientist
PDF, Word, Emails
Video Processing
NLP Engineer
Audio & Speech
Processing
Business Science



Top 10 profiles you can apply for after learning **Python**

01 Software Engineer

02 Web Developer/ Front-end Developer

03 Python Developer

04 DevOps Engineer

05 Research Analyst

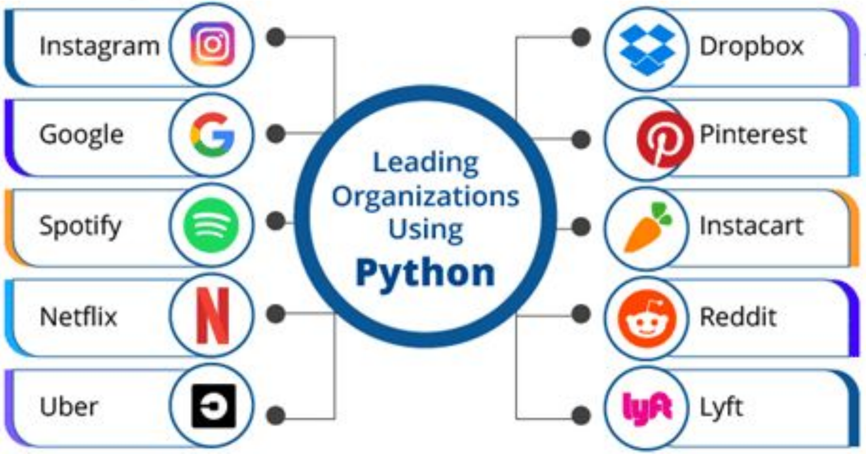
06 Data Analyst

07 Consultant

08 Product Manager

09 Data Scientist

10 Machine Learning Engineer

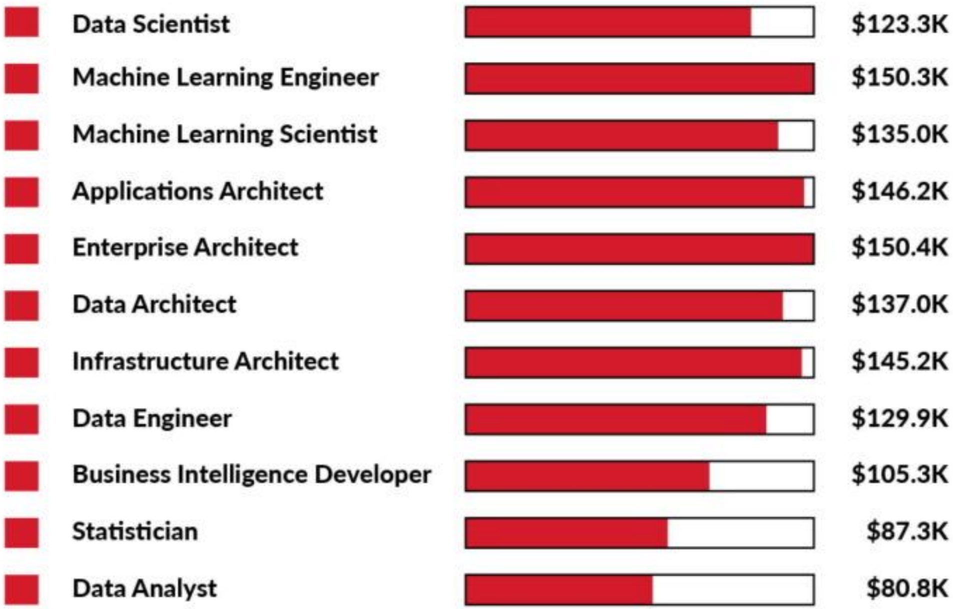


Leading Organizations Using **Python**

- Instagram
- Google
- Spotify
- Netflix
- Uber
- Dropbox
- Pinterest
- Instacart
- Reddit
- Lyft

Northeastern University

Salaries of In-Demand Data Science Jobs



Source: Lightcast™ Analyst, 2023

www.northeastern.edu

Another Perspective

The fastest growing companies in SV is either data or model companies: they operate on either big model or big models.

Fastest-growing data companies

snowflake

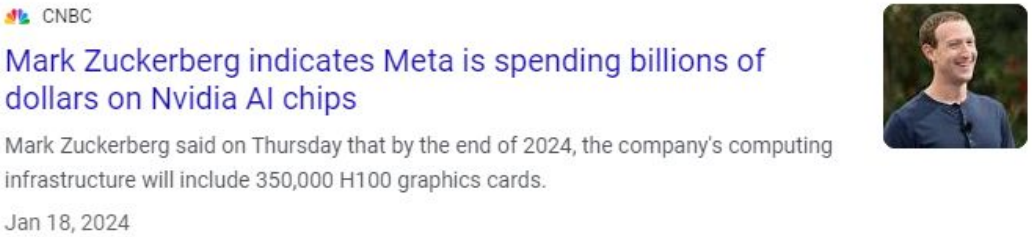
databricks

Fastest-growing model companies

OpenAI

ANTHROP\C

News

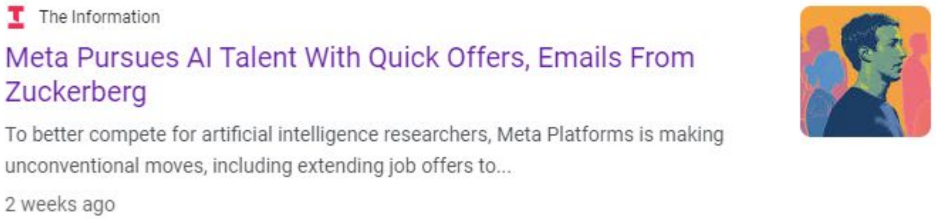


CNBC

Mark Zuckerberg indicates Meta is spending billions of dollars on Nvidia AI chips

Mark Zuckerberg said on Thursday that by the end of 2024, the company's computing infrastructure will include 350,000 H100 graphics cards.

Jan 18, 2024

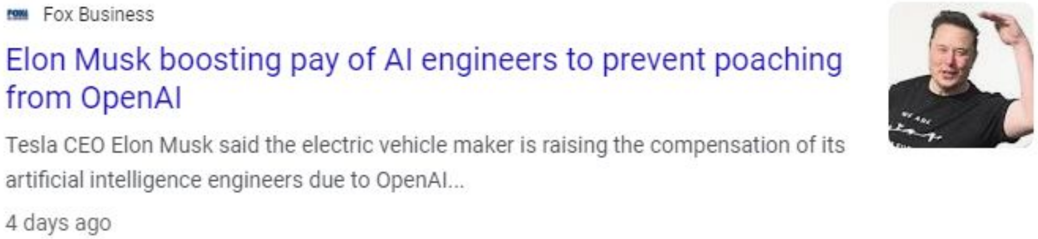


The Information

Meta Pursues AI Talent With Quick Offers, Emails From Zuckerberg

To better compete for artificial intelligence researchers, Meta Platforms is making unconventional moves, including extending job offers to...

2 weeks ago

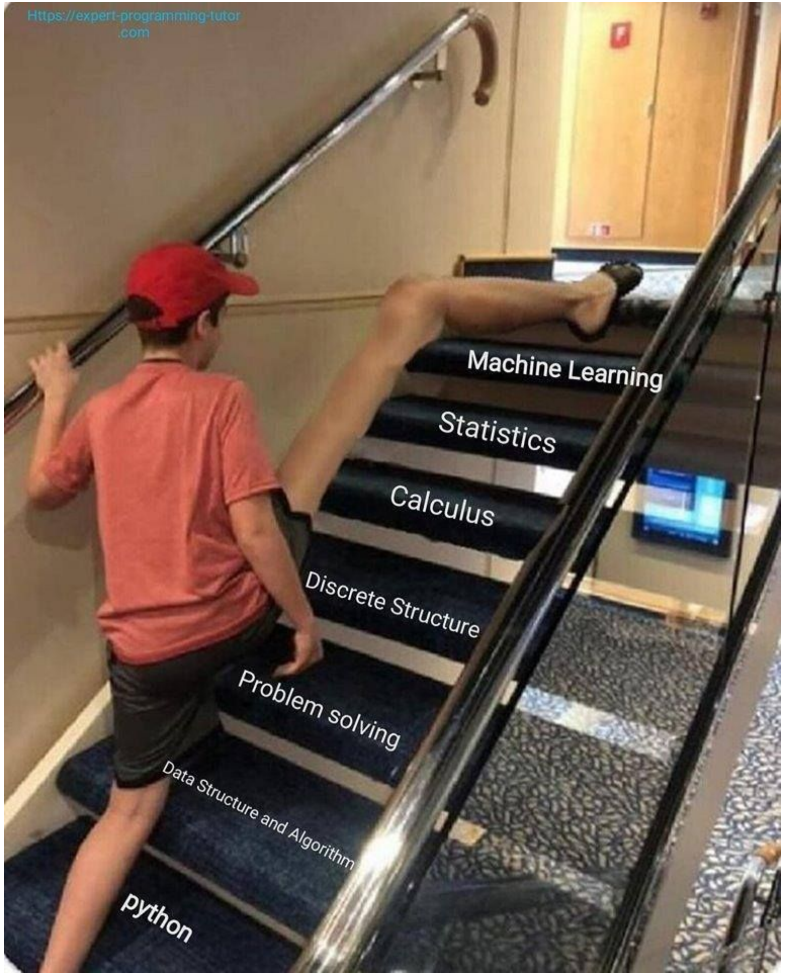


Fox Business

Elon Musk boosting pay of AI engineers to prevent poaching from OpenAI

Tesla CEO Elon Musk said the electric vehicle maker is raising the compensation of its artificial intelligence engineers due to OpenAI...

4 days ago


Https://expert-programming-tutor
.com
Machine Learning
Statistics
Calculus
Discrete Structure
Problem solving
Data Structure and Algorithm
python

Reading 1 (Part 1)
50 Years of Data Science

Overview:

- Connects the discipline of DS to its 50+ years of history (John Turkey in 1960s)
- DS as an extension of statistics?
- Common Task Framework (ex: Netflix Challenge)

DS as an extension of statistics?

- Multidisciplinary investigations (25%)
- Models and methods for data (20%)
- Computing with data (15%)
- Pedagogy (15%)
- Tool evaluation (5%)
- Theory (20%)

DS as an extension of statistics?

Inference model: to infer how nature is associating the response variables to the input variables.

Prediction model: to be able to predict what the responses are going to be future input variables

Professor Freiman’s paper is an important one for statisticians to read. He and Statistical Science should be applauded…His conclusions are consistent with how statistics is often practiced in business. -Bruce Hoadley

Common Task Framework and the secret sauce

- A publicly available training dataset
- A set of enrolled competitors
- A scoring referee

Common Task Framework and the secret sauce

1. Error rates decline by a fixed percentage each years to an asymptote depending on task and data quality.
2. Progress usually comes from many small improvements, a change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role- and is reused in unexpected ways.

Final Project Groups

Group Guide

Primer: https://kshannon-ucsd.github.io/cogs9/final-group-project/group-work-primer

- All group members are expected to contribute and be on the same page (chain is as strong as the weakest link)
- Random grouping to simulate real-work environment where you do not get to choose peers so please take it as a learning experience
- Each group should decide on a messaging channel (whatsapp, discord, google groups)
- We will communicate with each group individually for weekly check-in (through email or piazza group). Just to make sure that everything is going smoothly and everyone is contributing.
- In the check-in, each person can say what they did and we may ask any follow up questions if needed.

Announce Groups

Allow them to meet each other, pick team name, introduce and make messaging groups