COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

Discussion Section A01
Week 4

Teaching Assistant (TA): Abdullah
Instructional Assistant (IA): Kyra

Where to find all material

COGS 9

Search COGS 9 UCSD Podcast Gradescope

- Home
- Syllabus
- Readings
- Assignments
- Exam
- Final Project
- Office Hours
- Contact Us

This site uses Just the Docs, a

Introduction to Data Science

COGS 9 - UC San Diego - Prof. Kyle Shannon

Spring 2024 | SOLIS 107 | TU & TH 5:00-6:20PM

Welcome 👋

We are all very excited that you decided to join us on this whirlwind tour of data science. All relevant info, e.g. due dates, assignment links, etc. are found on this website. We look forward to teaching and working with all of you and hope to meet you in office hours. Check out the **Getting Started** section so you can hit the ground running when class starts!

NOTE
Week one I try to take as many students from the **waitlist** as I can, please email cogsadvising@ucsd.edu with further questions.

Discussion Sections

	Day	Time	Location	Staff	Materials
A01	Wed	12:00-12:50PM	CENTR 222	TA: Abdullah IAs: Kyra	View
A02	Wed	1:00-1:50PM	CENTR 222	TA: Kaushik IAs: Seshu, Vicky	View
A03	Wed	2:00-2:50PM	CENTR 222	TA: Matthew IAs: Jessica, Wenhua	View
A04	Wed	3:00-3:50PM	CENTR 222	TA: Vineeth IAs: Jiesen	View
A05	Wed	4:00-4:50PM	CENTR 222	TA: Vineeth IAs: Harshita	View

cogs9_TA Public

main | 1 Branch | 0 Tags | Go to file

AbdullahAshfaq Added week3 material

week2	Added week2 material
week3	Added week3 material
README.md	Update README.md

README

Cogs 9 Discussions-Intro to Data Science

Abdullah's discussion section material for COGS9 course

Upcoming Deadlines

Week 4		
Tue, Apr 23	LECT	Data Wrangling
Thu, Apr 25	LECT	Programming for Data Science
Fri, Apr 26	QUIZ	Reading Quiz 2 due
Fri, Apr 26	READ	Begin reading 3

Week 5		
Tue, Apr 30	LECT	Data Viz & Descriptive Analysis
Thu, May 02	LECT	Exploratory Data Analysis
Fri, May 03	QUIZ	Reading Quiz 3 due
Fri, May 03	READ	Begin reading 4

Week 6		
Mon, May 06	ASSG	Assignment 1 due
Tue, May 07	LECT	Communicating Data Science
Thu, May 09	LECT	Inferential Analysis

Week 7		
Mon, May 13	PROJ	Final Project Part 1 due

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 SQL
- 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

Today’s Outline

Participation = Extra Credit 😀

- Reading 2
- Getting Data
  - Downloading
  - API
  - web scraping
- Data Wrangling in Python
  - Subset dataset
  - Change order
  - Add or modify column
  - Summarize data

Reading 2
Ethics, Privacy, Security

What is “PII”

“Any information relation to an … natural person … who can be identified, directly or indirectly, in particular by reference … to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity.”

- Data Protection Detective

PII Privacy Protection Technologies

Examples: k-anonymity, l-diversity

“These methods aim to make joins with external datasets harder by anonymizing the identifying attributes.”

The methods modifies quasi-identifiers to satisfy various syntactic properties to prevent.

Problem: Simply not enough to do the job

Re-identification without PII

It turns out there's a wide spectrum of human characteristics that enable re identification as long as they satisfy the following key properties:

1. They are reasonably stable across time and contexts
2. The corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability.

Re-identification algorithms take advantage of such properties to re-identify individuals using other attributes

Differential Privacy

Differential privacy (DP) is a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.

Go to Notebooks