

Introduction

Data Science Tools 1

Fall 2021



UNIVERSITY *of*
DENVER

Agenda

- Introduction
- Housekeeping
- Review Wk 1
- Review HW 1
- Questions
- Additional Topics

- Name
- Geographic Location
- Academic background
 - Undergrad/grad
 - Are you in Data Science Program? If not, which program are you enrolled in? How far along are you in in your program?
- Professional experience (optional)

- Name
- Geographic Location
- Academic background
 - What was your undergrad in?
 - How far along are you in this program?
- Professional experience (optional)

- Diwanshu Shekhar
(Dee-'Waan-Shoo)
- PhD Candidate in Computer Science
- 12 years of experience in the industry
 - Software Engineering
 - Data Engineering
 - Machine Learning
- Blog: hiddenlayer.net

- Async Lectures *Soft* Deadline
 - Before Live Session begins
- Assignments Due Date
 - Every Sunday by Midnight
- Mid-term and assignment due the same week
- Final project
- ~~https://git.es.du.edu/psnegi/online_tools1_notebooks~~
- https://github.com/psnegi/data_science_tools1
- Office Hours
 - Every Friday from 3-5 pm MST
 - Online (calendar in Atrio) or AAC (Anderson Academic Commons)
- Communication
 - CS Department Slack Channel - compscidu.slack.com
 - Email: diwanshu.shekhar@du.edu
 - Phone: +1- 657-888-3492 (for urgent and quick matters only)

- What is expected of students before live session?
 - Go through async materials
 - Go through the next homework assignment
 - Answer async discussions
- What students can expect from this class?
 - Covers breadth of materials
 - This is not a specialization class - we don't go deep into any particular subject area



- Why and how of Python Virtual Environment
- Jupyter Notebook
- Brief Intro to Data Science
- Brief Introduction to Git

- Why?
 - Isolation
- Other virtual environments
 - Conda virtual environment
 - Poetry

- Pros
 - Markdown, HTML, Latex
 - Executable Cell
- Cons
 - Not for software engineering
 - Cannot write unit tests
 - Cannot use one notebook into another
 - Cannot make your code modular
 - Order of execution in notebook is prone to cause errors

- Data collection
- Data cleaning
- Exploratory Data Analysis (EDA)
 - Visualization
- Machine Learning
 - Model selection
 - Model evaluation

- De-centralized version control system
- Git Workflow
 - `git add`
 - `git commit`
 - `git push`
 - `git branch`
 - `git rebase/merge`

Homework 1

- Review
- Due 11:59 PM Sunday



- <https://docs.anaconda.com/anaconda/user-guide/getting-started/>
- Anaconda Navigator
- Command Line

- <https://code.visualstudio.com/>

- <https://www.jetbrains.com/pycharm/>

