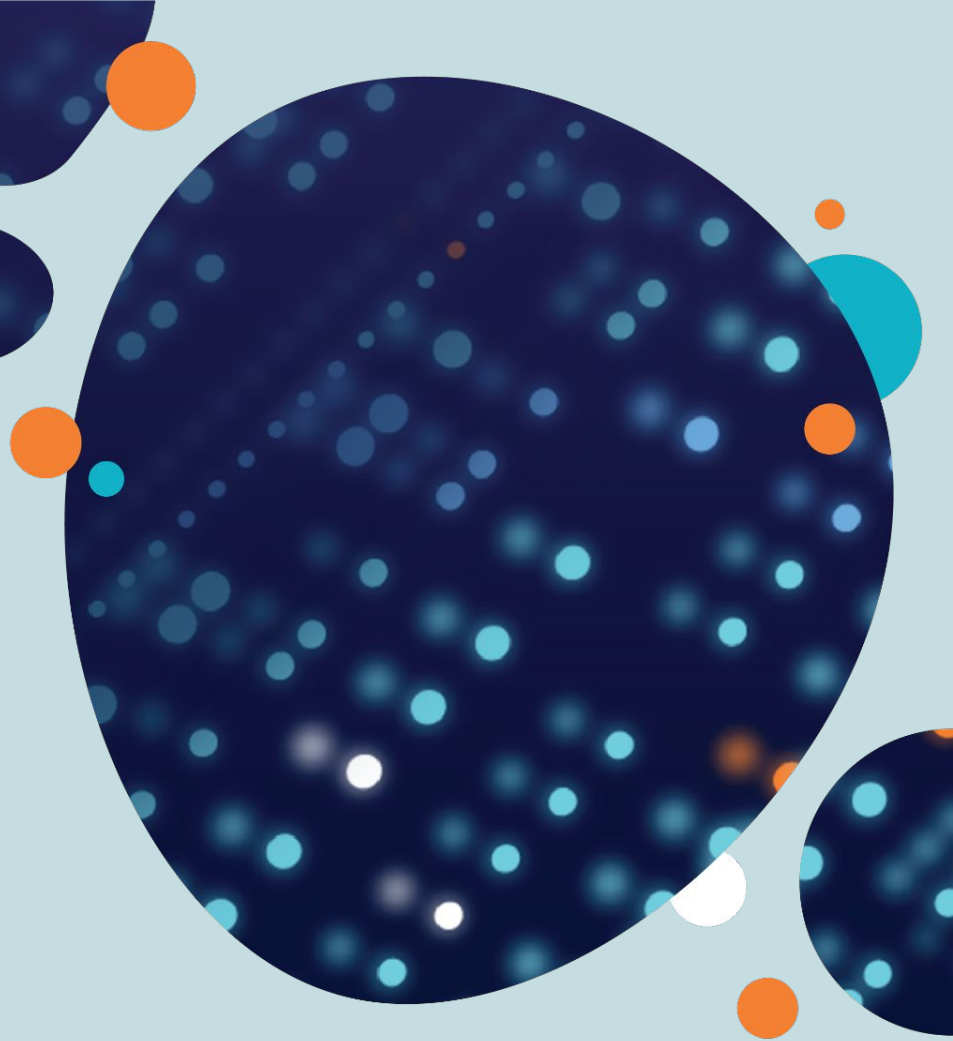
data science
student society

Introduction to Data Science Projects

October 10, 2024



data science student society

Please check in!

What is Data Science?

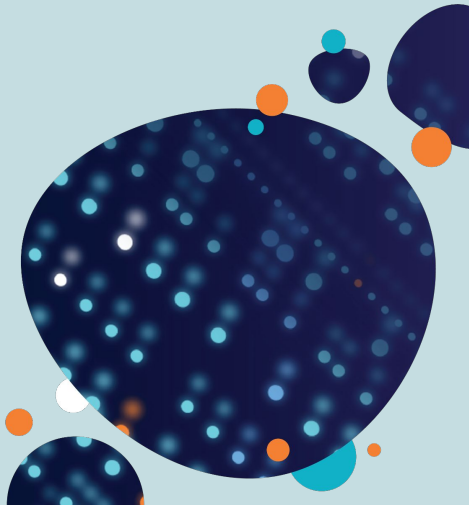
data science student society

- While there is no one true definition of data science, generally:
- Combination of statistics and computer science designed to analyze large sets of data
- Usually uses machine learning and AI
- Requires high levels of research and specialized domain knowledge to draw conclusions

Website: ds3ucsd.com

Data Science Project Essentials

- Version Control: Git and Github
- IDE
- Programming Language and Virtual Environment

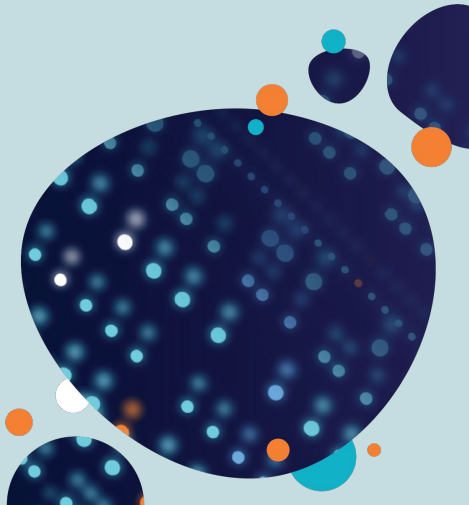


Website: ds3ucsd.com

Version Control: Git and Github

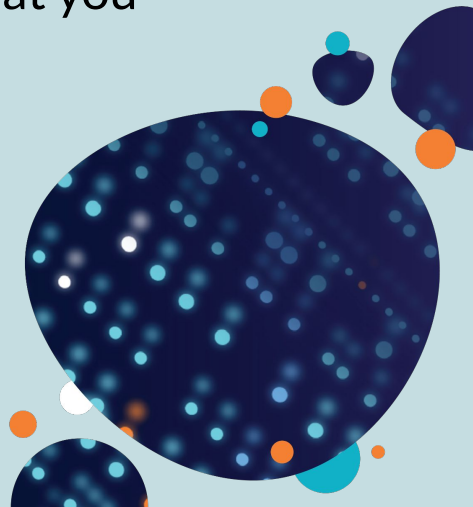
- Git is a version-control scheme
  - A “save state” of sorts for your projects
  - Allows you to revert to previous versions of your project
  - Also allows for collaboration on the same project
- Github is a repository for Git
- Basic Git commands:
  - `git clone path/to/repository`
  - `git commit -m`
  - `git push`

More commands here: https://confluence.atlassian.com/bitbucketserver/basic-git-commands-776639767.html



What's an IDE?

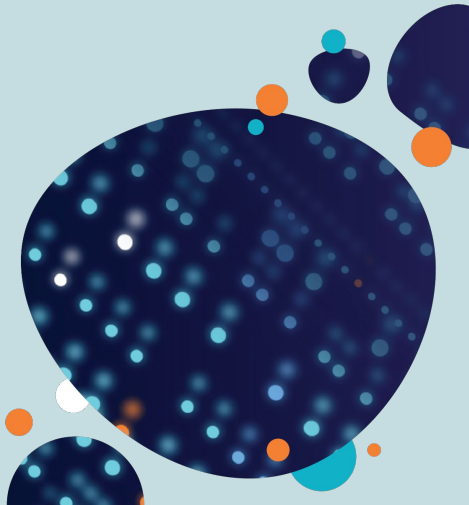
- Stands for "Integrated Development Environment"
- Is a place where you can create, run, and test your code
    - Often has convenient shortcuts/tools that you can use for your projects
- The one you use depends on your preferences and what you plan to work on.
- Examples include Visual Studio Code, Android Studio



Virtual Environments (venv)

- Virtual environments are virtual spaces separate from your installed programs
- This allows you to avoid messing with libraries with other projects
    - Also allows you to work on a project with different versions of the same language, e.g. Python 10.3 vs. Python 10.4
- Examples: conda

Create a venv with pip:
https://www.geeksforgeeks.org/python-virtual-environment/



Data Science VS Code Setup

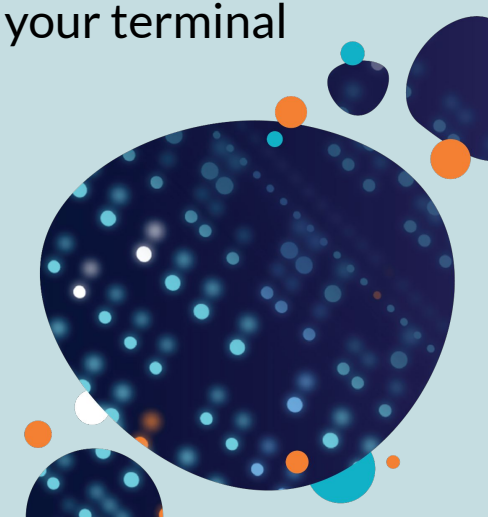
- Install VS Code here: https://code.visualstudio.com/
- Install Python, checking the box Add Python to PATH: https://www.python.org/downloads/
- Install Anaconda, add it to system PATH: https://www.anaconda.com/download
- Install Python extension in VS Code
- In VS Code terminal:
  - Install Jupyter by running `conda install jupyter`
  - Create a conda environment by running `conda create -n myenv python=3.14`
- Install Jupyter extension in VS Code

More commands here: https://confluence.atlassian.com/bitbucketserver/basic-git-commands-776639767.html

Cloning the Notebook

- Create a Github account: https://github.com/
- Download Git: https://git-scm.com/downloads
- Next, download Github desktop: https://github.com/apps/desktop
- Clone the repository here in your terminal: https://github.com/couchsnail/ds3-workshops
- Open the Notebook and run `conda activate myenv` in your terminal
- Select `myenv` as the Kernel for the Notebook
- The Notebook is a simplified version of how you'd start a data science project. Try out some of the problems for yourself!

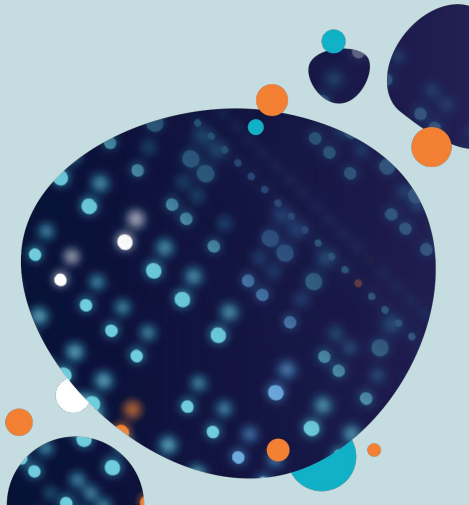
Note: If it says pandas or numpy doesn't exist, run
`conda install pandas numpy`



Further Resources

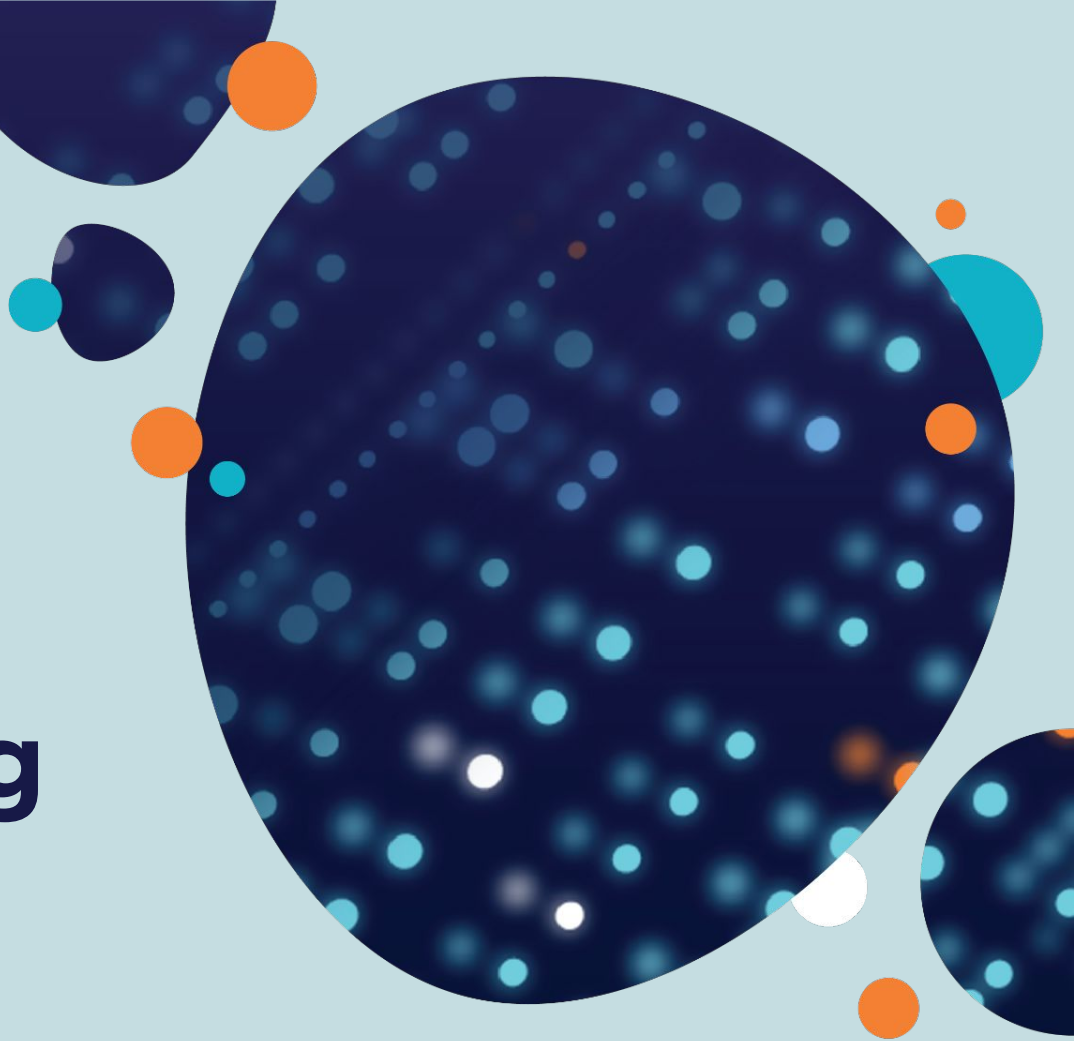
- A more in-depth tutorial to Git and Github: https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
- Github and VS Code tutorial: https://vscode.github.com/
- How to use Jupyter Notebooks on VS Code: https://code.visualstudio.com/docs/datascience/data-science-tutorial

Note: Make sure to stay connected on the DS3 Discord for further workshops!!!!



data science
student society

Next Time: Intro to Web-Scraping



data science student society

Leave your feedback here!