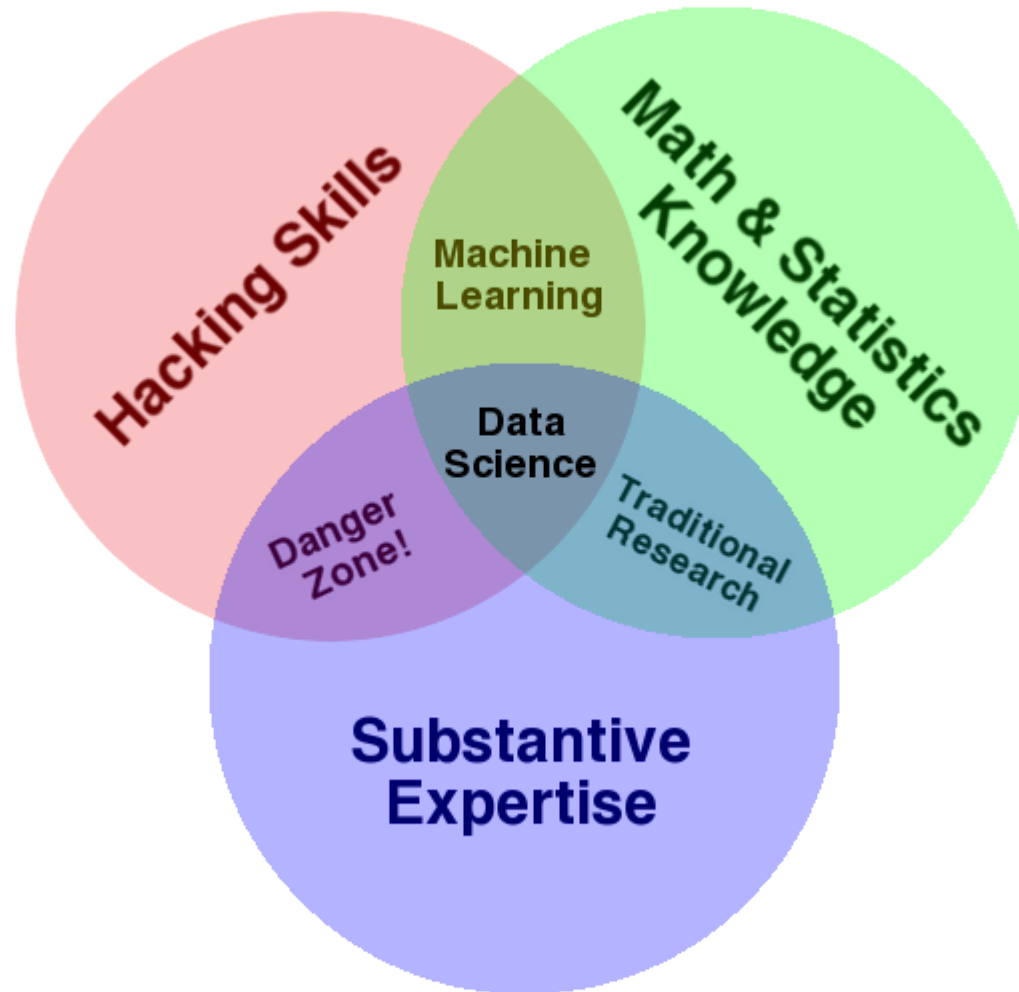# Data Science: Intro

**CPSC 501: Advanced Programming Techniques**
**Fall 2022**

Jonathan Hudson, Ph.D
Assistant Professor (Teaching)
Department of Computer Science
University of Calgary

**Monday, October 3, 2022**

# What is Data Science?

2

UNIVERSITY OF CALGARY

# Python

- **Python** 3 – Most popular Data Science programming language
  - **R** is popular from the Statistics side of things, **Julia** in Mathematics

1. Very strong support of packages, tutorials, and knowledge
2. Ease of integrating more efficient languages behind scenes like **C++**, etc.
3. Often prototype in **R**/others but implement in **Python** for final production

1. Interpreted, so can be slow (unless break out to **C++** like with *numpy*)
2. Not built for multi-thread concurrency without effort (unless break out like with *tensorflow* for neural networks to use multi-core/GPU)

UNIVERSITY OF
CALGARY

# Data Science Installation Method (managed)

- **Anaconda** – Distribution of **Python** 3 and package manager that allows easy access to most popular data science libraries

- **Miniconda** – Lighter weight than **Anaconda** as it doesn't pre-download as many packages
  - Generally early learning path is to get miniconda
  - https://docs.conda.io/en/latest/miniconda.html
  - Then install

```
1 conda install numpy pandas matplotlib seaborn ipython jupyterlab
2 conda install scipy scikit-learn tensorflow keras statsmodels
```

CONDA    MINICONDA    ANACONDA

= conda
+Python.exe
+base modules

= miniconda
+150 modules
+ user interface
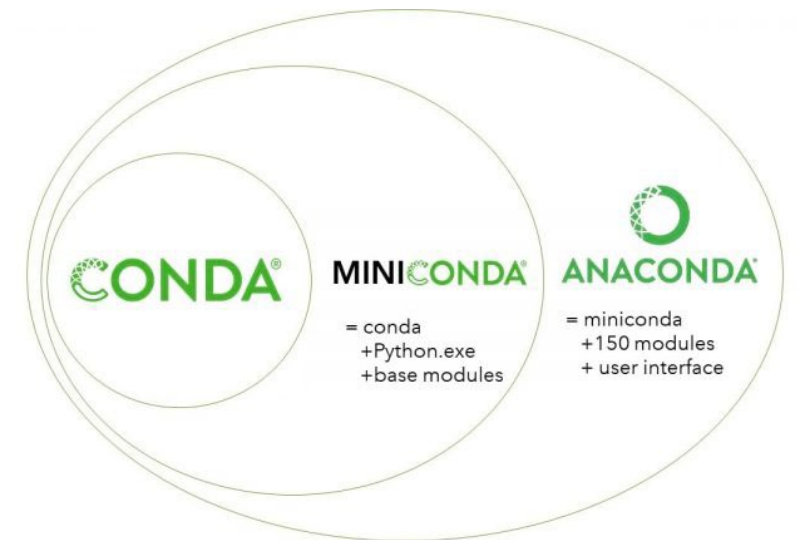
The first is for Data Science tasks while the second line of installation is for Machine Learning

UNIVERSITY OF CALGARY

# Data Science Installation (lightweight)

- Could just use available **Jupyter** notebook *ipython*-like environments such as

1. **Google Colaboratory** https://colab.research.google.com/ (public)

2. **Syzygy** https://ucalgary.syzygy.ca/ (university member host)

- Or setup your local install of **Python**
  - IDE choice up to you (for Python prefer I like **Pycharm**)
  - Recommend virtual environment (good for UofC lab machines)
  - **pip** for packages (python package manager)

UNIVERSITY OF CALGARY

# Data Science Installation (IDE-Pycharm)

- Bottom bar of **Pycharm**



- Can use terminal to install packages to main IDE **Python** install
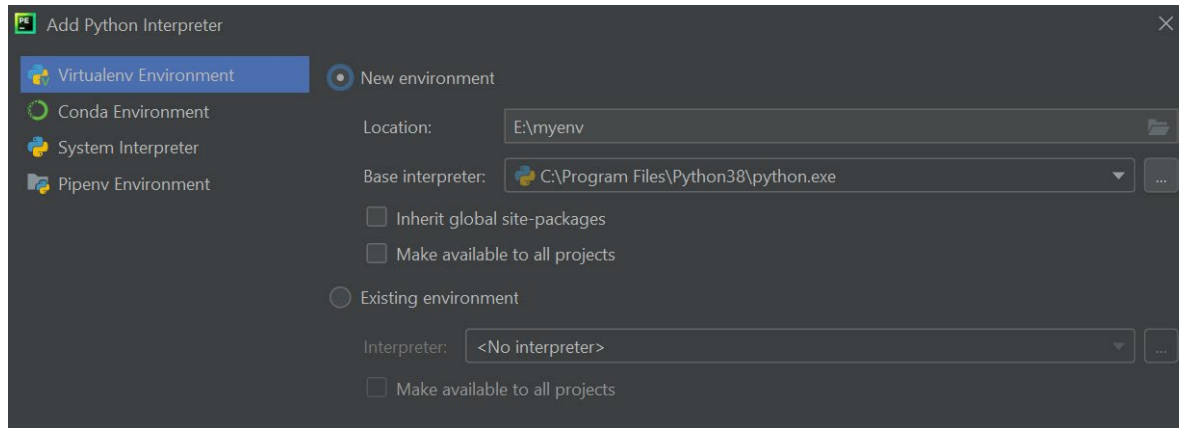


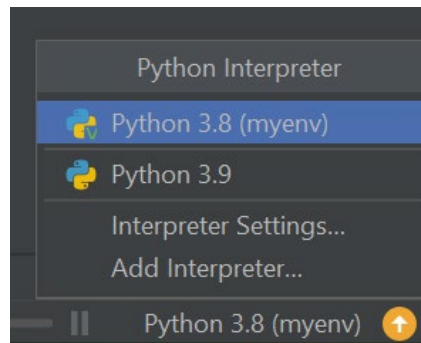- Can also use package manager to do same
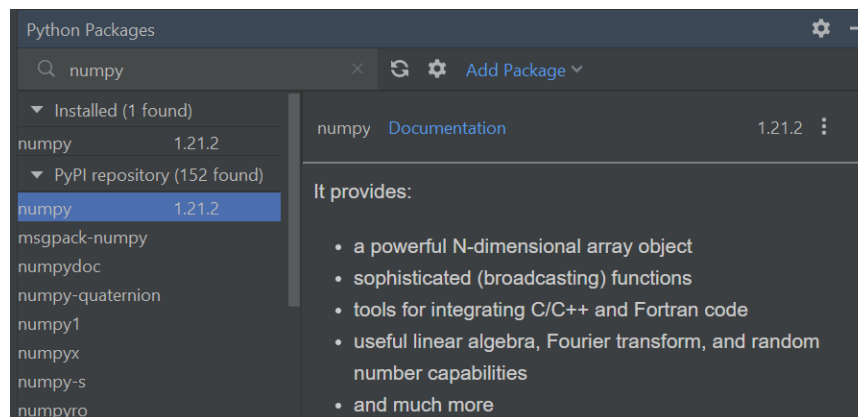
# Data Science Installation (IDE-Pycharm)

- Instead of using system interpreter you can create virtual environment



- Bottom right corner you can toggled between **Python** interpreters

# Data Science Installation (windows shell)

```
1 python -m venv myenv
2
3 .\myenv\Scripts\activate
4
5 pip install numpy
6 pip install pandas
7 pip install matplotlib
8 pip install seaborn
9 pip install ipython jupyterlab notebook
10
11 pip install scipy scikit-learn tensorflow keras statsmodels
12
13 deactivate
```

UNIVERSITY OF
CALGARY

# Data Science Installation (unix shell)

```
 1 python -m venv myenv
 2
 3 source myenv/bin/activate
 4
 5 pip install numpy
 6 pip install pandas
 7 pip install matplotlib
 8 pip install seaborn
 9 pip install ipython jupyterlab notebook
10
11 pip install scipy scikit-learn tensorflow keras statsmodels
12
13 deactivate
```

UNIVERSITY OF
CALGARY

# Data Science Installation (docker)

- Examples of data science oriented docker containers
  - pytorch/pytorch — a simple container for Use Case 1 that includes Pytorch
  - jupyter/scipy-notebook — A container for Use Case 2 that includes Jupyter as the UI, and many python data science modules.
  - DAGsHub/ml-workspace-minimal —This container is an updated version from the ml-tooling/ml-workspace repository. The original has not been maintained for the last !9 months so this is updated version.
    - - Jupyter, JupyterLab, VSCode web-based IDE, Pytorch, Tensorflow, Sklearn, Pandas, and many other popular data science libraries & tools.
    - - 🗎 Full Linux desktop GUI accessible via a web browser, Easy terminal access via a web browser,  Seamless Git integration optimized for notebooks
    - Integrated hardware & training monitoring via Tensorboard & Netdata.
    - Access from anywhere via Web, SSH, or VNC under a single port, Usable as a remote kernel (Jupyter) or remote machine (VSCode) via SSH
    - Easy to deploy on Mac, Linux, and Windows via Docker.

UNIVERSITY OF CALGARY

# Packages - Interaction

**IPython** (2001)



- https://ipython.org/

- This is a interactive version of **Python**. Differences from regular interactive is that it can store state for reference using In and Out blocks (stored as arrays), and you can also break out of it to run shell commands from inside (like to install library or make folders and manage files)

**jupyter** notebooks (2014)



- https://jupyter.org/

- GUI to **IPython**,. Your computer hosts a execution kernel and you access it using a web browser web-based interface (like the **Google Colaboratory** and **Syzygy** options given earlier) Can do other languages as well.

# Packages – Standard Data Science

- *numpy* – manipulation of homogenous array-based data, container
  - https://numpy.org/



- *pandas* – 2010, heterogenous and labeled data like tables or databases, brings spreadsheets like functionality to data handling (on top of *numpy*)
  - https://pandas.pydata.org/

- *matplotlib* – publication quality visuals
  - https://matplotlib.org/



- *seaborn* – better charts on top of *matplotlib* (compete with **R**)
  - https://seaborn.pydata.org/

UNIVERSITY OF CALGARY

# Packages – Machine Learning

- *scipy* – *scikit* uses for algorithms, good at storing sparse, common scientific computing tasks (integration, linear algebra, optimization, signal processing, statistics distributions/variables/etc.)
  - https://scipy.org/

- *scikit-learn* – 2010 machine learning, classification, regression, clustering, etc.
  - https://scikit-learn.org/stable/index.html

- *tensorflow* – 2015 open-sourced distributed neural network library, 2019 v2.0
  - https://www.tensorflow.org/

- *keras* – 2015 deep learning library (often on top of *tensorflow*)
  - https://keras.io/

- *statmodels* – classical statistics and econometrics
  - https://www.statsmodels.org/stable/index.html

# Importing Libraries in Python (style)



```python
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 import tensorflow as tf
6 import statsmodels as sm
```

UNIVERSITY OF
CALGARY

# Onward to ...
# IPython.

Jonathan Hudson
jwhudson@ucalgary.ca
https://pages.cpsc.ucalgary.ca/~jwhudson/

UNIVERSITY OF
CALGARY