DATA SCIENCE

COURSE INSTRUCTOR: MUHAMMAD SAIF UL ISLAM

Lecture Outline – Week#1

➢ Introductory words

➢ Introduction to the Course

➢ Discussion on Course outline

➢ Course plan, Assignments and Project

➢ Introduction to Data Science

➢ Applications of AI & Data Science

➢ Characteristics of Data Scientist

➢ Installing Python/Anaconda, Agent & Environment

About Myself
Muhammad Saif ul Islam

Education:

PhD Scholar (Computer Science)

➢ FAST-NUCES, LHR

Masters in Data Science - 2019

➢ FAST-NUCES, KHI

Bachelors in Computer Science -2017

➢ Bahria University, KHI

Work Experience:

IT Instructor – 5 Months

➢ IBA-BBSYDP

Python Developer – 7 Months

➢ Innovative Solutions

Sr. Operations Engineer – 1 Year

➢ Gfk Etilize

Lecturer – 2.5 years

➢Mohammad Ali Jinnah University

Lecturer – 6 Months

➢Beaconhouse National University

Lecturer – Since Spring 2023
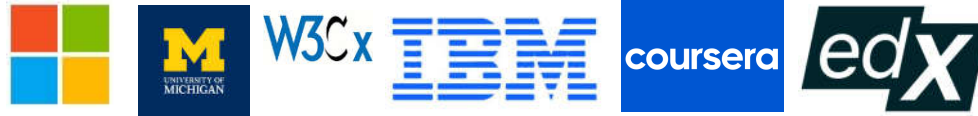
➢ FAST- NUCES

About Myself
Certifications

Python:
- DAT210x: Programming with Python for Data Science
- Introduction to Python for Data Science
- Introduction to Data Science in Python
- Python for Everybody
- Python Data Structures

Database:
- Using Databases with Python
- Querying Data with Transact-SQL

Data Science:
- Data Science Essentials
- Python Project for Data Science
- Applied Plotting, Charting & Data Representation in Python
- Capstone: Retrieving, Processing, and Visualizing Data with Python
- Applied Machine Learning in Python
- Image Processing with Python

Web:
- Using Python to Access Web Data
- HTML5 Introduction

About Myself
Publications

Mustafa Khan, M., **Ul Islam, M. S.,** Siddiqui, A. A., & Qadri, M. T. (2023). Dual deterministic model based on deep neural network for the classification of pneumonia. *Intelligent Decision Technologies*, *17*(3), 641–654. https://doi.org/10.3233/idt-220192

Muhammad Saif ul Islam, Using deep learning based methods to classify salt bodies in seismic images, Journal of Applied Geophysics, Volume 178, 2020, 104054, ISSN 0926-9851, https://doi.org/10.1016/j.jappgeo.2020.104054.

M. Mehboob, M. S. Ali, **S. Ul Islam** and S. Sarmad Ali, "Evaluating Automatic CV Shortlisting Tool For Job Recruitment Based On Machine Learning Techniques," 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2022, pp. 1-4, doi: 10.1109/MAJICC56935.2022.9994112.

M. S. ul Islam and H. Farooq, "Rating visual contents of website using brain computer interface," 2017 International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, 2017, pp. 23-27, doi: 10.1109/ICICT.2017.8320159.

Students' Introduction

Expectation:

➢ What did you choose this course?

➢ What do you want to learn in this course?

Name?

Course Outline

➤ Introduction to Data Science and Applications

➤ Overview of Data Science Process

➤ Introduction to Python Programming
- Operators, Loops, Conditions, List, Functions, File Handling

➤ **Data Retrieval**
- Numpy basics and Pandas
- Acquiring Data
- Web Scrapping

➤ **Data Preparation**
- Data Wrangling and Transformation
- Indexing, Selecting and Slicing Data
- Working with duplicate, missing data and outliers
- Textual Categorical Features and Conversion

➤ **Data Exploration and Visualization**
- Distributions, Variance, and Standard Deviation
- Covariance and Correlation
- Working with matplotlib
- Histogram, Bar Chart, Scatter Plot, Pie Plot, 3D Plot, Box plots, Heat Maps etc.
- Curse of Dimensionality and Dimensionality reduction with Principal Component Analysis

➤ **Data Modeling and Evaluation**
- Linear Regression and Interpretation of coefficients
- Regression Analysis: Case study
- R-Squared; Coefficient of determination, Mean Squared Error
- Model Evaluation and Confusion Matrix
- K-Mean Clustering
- Logistic Regression

Course plan, Assignments and Quizzes

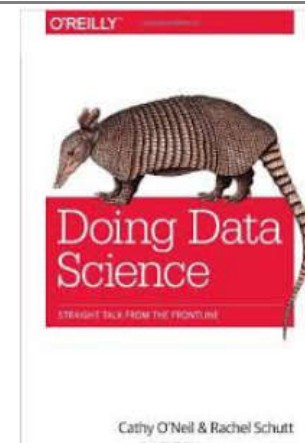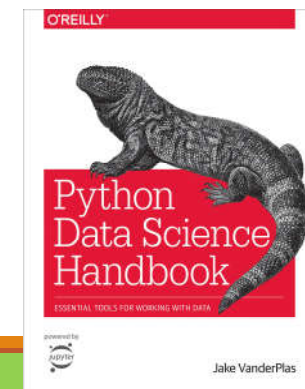|	Graded Assessment types	Weights (%)
1	Project + Research paper	15%
2	Quiz	5%
3	Assignments	10%
4	Mid Exam	30%
5	Final Assessment	40%
	Total:	100%

Resources

Books:

Doing Data Science by Oreilly

Python Data Science Handbook: Essential Tools for Working with Data Book by Jake VanderPlas

Consulting Hours

Contact at:

Email: **saif.islam@nu.edu.pk**

Office Hours:

**Room# MLO-007**

Mon, Wed 3:00PM-4:00PM

OFF Days: Sat, Sun

What is Data Science?

➢Data Science is the process of slicing through massive chunks of data, processing and analyzing them for meaningful information that can help businesses get insights on concerns, customer experience, supply-chain and other prime aspects that would complement their business operations.

➢Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data

➢Data science principles apply to all data – big and small

Data Science is Multidisciplinary
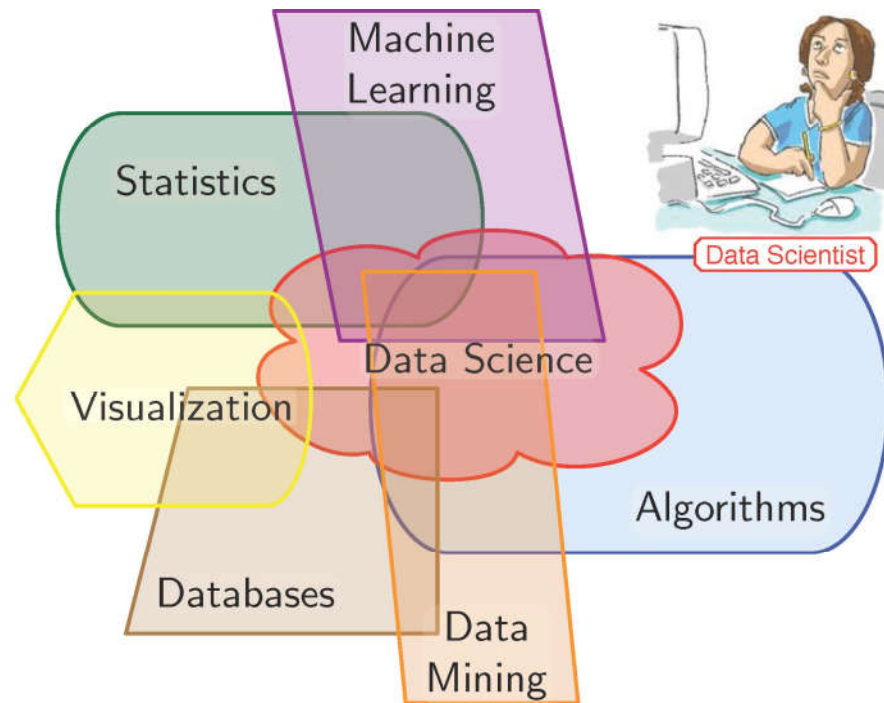
Who is Data Scientist?

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and **decisions**

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Miscellaneous

Kaggle.com
◦ The Home of Data Science
◦ Company recruiting & Pays winners
◦ Many Kaggle winners manage Analytics teams
◦ Compete! Get recognized.
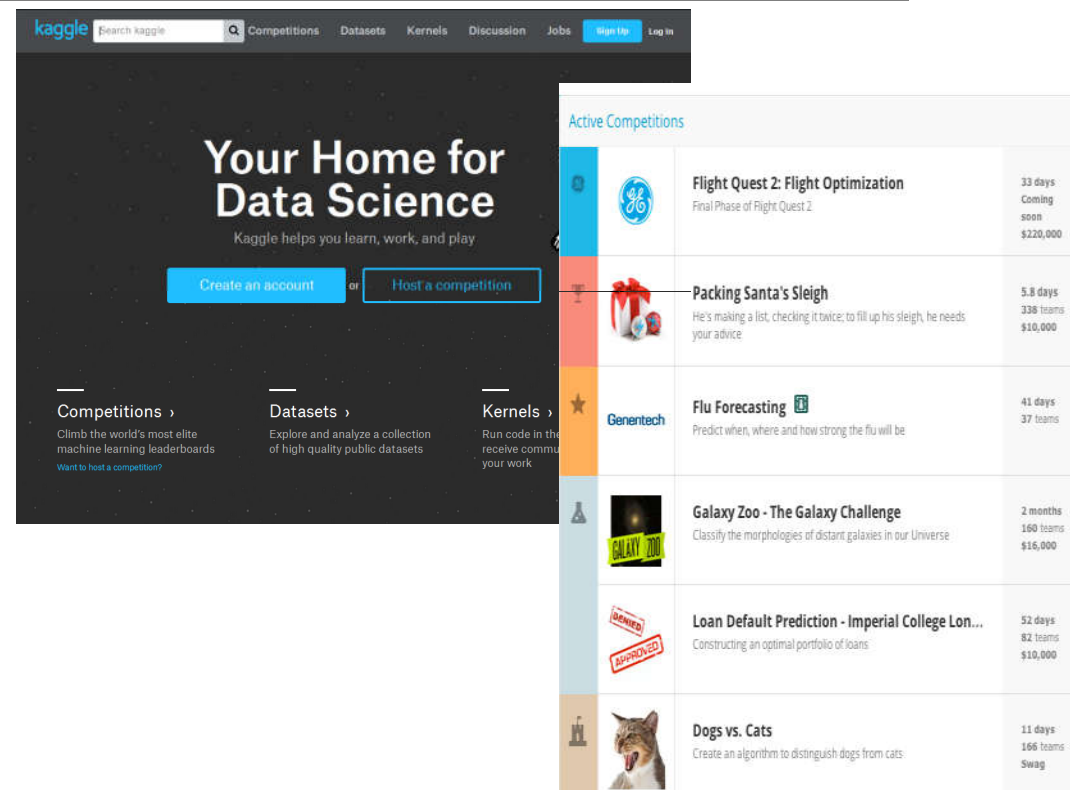
Real Life Applications of AI & Data Science

➢ Marketing

➢ Finance

➢ Agriculture

➢ HealthCare

➢ Gaming

➢ Space Exploration

➢ Autonomous Vehicles

➢ Artificial Creativity

Real Life Applications of AI & Data Science
Marketing

AI generated content: An AI writing program called 'WordSmith' produced 1.5 billion pieces of content in 2016, and is expected to grow further in popularity in the coming years.

Smart Content Curation: Allows you to better engage visitors on your site by showing them content relevant to them. Cross selling, personalized messaging, recommendation etc.

Smart Search: Search engines read our minds and provide all possible results related to the item, Voice-search technology (Google, Amazon, Apple), Interpret consumer's queries -Chatbots.

Predictive analytics: Predicting the likelihood of a given customer to convert, predicting what price a customer is likely to convert at, or what customers are most likely to make repeat purchases. Propensity modeling.

Dynamic pricing: Dynamic pricing can nudge interested consumers into becoming customers by targeting only special offers only at those likely to need them in order to convert.

Real Life Applications of AI & Data Science
Banking & Finance

Recommendation Engines: In the banking sector, the system learns from the user's behavior. Based on the previous actions, it can recommend appropriate investment strategies, credit card plans, and make other offers that would save the user a lot of time browsing through the website.

Fraud Detection and Prevention: Based on self-learning artificial technology and real-time behavioral profiling, the system can detect suspicious behavior and prevent frauds.

Trading: Investment companies have been relying on computers and data scientists to determine future patterns in the market. As a domain, trading and investments depend on the ability to predict the future accurately.

Predictive analytics: Uses real-time and historical data to deliver precise information that helps traders to quote a better price when selling and buying bonds for their clients.

Real Life Applications of AI & Data Science
Agriculture

Forecasted Weather data: The forecasted/ predicted data help farmers increase yields and profits without risking the crop. By implementing such practice helps to make a smart decision on time.

Monitoring Crop and Soil Health: Utilizing AI is an efficient way to conduct, or monitor identifies possible defects and nutrient deficiencies in the soil. With the image recognition approach, AI identifies possible defects through images captured by the camera.

Decrease pesticide usage: With the help of the AI, data are gathered to keep a check on the weed which helps the farmers to spray chemicals only where the weeds are. This directly reduced the usage of the chemical spraying an entire field.

AI Agriculture Bots: AI bots in the agriculture field can harvest crops at a higher volume and faster pace than human laborers. By leveraging computer vision helps to monitor the weed and spray them.

Real Life Applications of AI & Data Science
Health Care

Medical Imaging: With AI in medical imaging, treatments can be personalized, and results can be transmitted with ease. Doctors can also efficiently identify cardiovascular disorders along with other fractures and injuries. Cancer cells detection, brain tumor detection, pneumonia detection etc. are few example.

Robot Assisted Surgery: In orthopedic surgery, a form of AI-assisted robotics can analyze data from pre-op medical records to physically guide the surgeon's instrument in real-time during a procedure. It can also use data from actual surgical experiences to inform new surgical techniques.

Automated Diagnosis and Error Reduction: In 2017, a group at Stanford University tested an AI algorithm against 21 dermatologists on its ability to identify skin cancers. The clinical findings, as reported by Nature last year, "artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists."

Virtual Nurses: To interact with patients, ask them questions about their health, assess their symptoms, and direct them to the most effective care setting. Molly, etc.

Real Life Applications of AI & Data Science
Gaming

AlphaGo: DeepMind's AlphaGo is the first computer program to defeat a professional human Go player (GrandMaster)

AlphaZero: AI beats champion chess program 'StockFish' after teaching itself in four hours.

Intelligent behaviors in characters: In video **games**, **artificial intelligence** (**AI**) is used to generate responsive, adaptive or **intelligent** behaviors primarily in non-player characters (NPCs) similar to human-like **intelligence**

Adversarial searches: Examples are Chess, Checkers, Go, etc.

Real Life Applications of AI & Data Science
Space Exploration

Spacecraft Monitoring and Control: Machine learning algorithms have been used in monitoring the spacecraft, autonomous navigation of the spacecraft, controlling systems, and intelligently detecting objects in the route

AI Based Assistants: AI-based assistants are being created to aid astronauts in their missions to Mars and beyond. These assistants are designed to understand and predicts the requirements of the crew and comprehend astronauts' emotions and their mental health.

Space Imaging and Exploration: According to the European Space Agency (ESA), satellites can produce over 150 terabytes of data per day. With the use of AI technologies, one can reduce the mission costs, extend battery life, and can analyze a vast amount of imaging data produced by the satellites. Example: Earth Observer 1 (EO-1) satellite, SKICAT, ENVISAT etc.
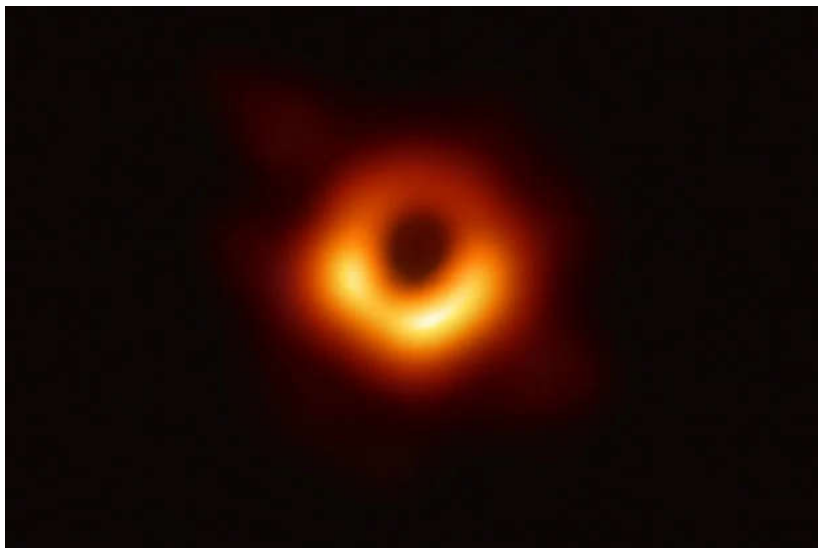
With the help of Google's trained model, NASA also managed to discover two obscure planets — **Kepler-90i and Kepler-80g**.

The creation of the algorithm that made the **first black hole image** possible was led by MIT grad student **Katie Bouman**

Real Life Applications of AI & Data Science
Space Exploration (Continue..)

Real Life Applications of AI & Data Science
Autonomous Vehicles

Waymo: n April 2017, Waymo started a limited trial of a self-driving taxi service in Phoenix, Arizona. On December 5, 2018, the service launched a commercial self-driving car service called "Waymo One"; users in the Phoenix metropolitan area use an app to request a pick-up

Advanced Driver Assistance Systems (ADAS): Camera-based machine vision systems, radar-based detection units, driver condition evaluation and sensor fusion engine control units (ECUs).

Infotainment human-machine interface: Speech recognition and gesture recognition, eye tracking and driver monitoring, virtual assistance and natural language interfaces.

Real Life Applications of AI & Data Science
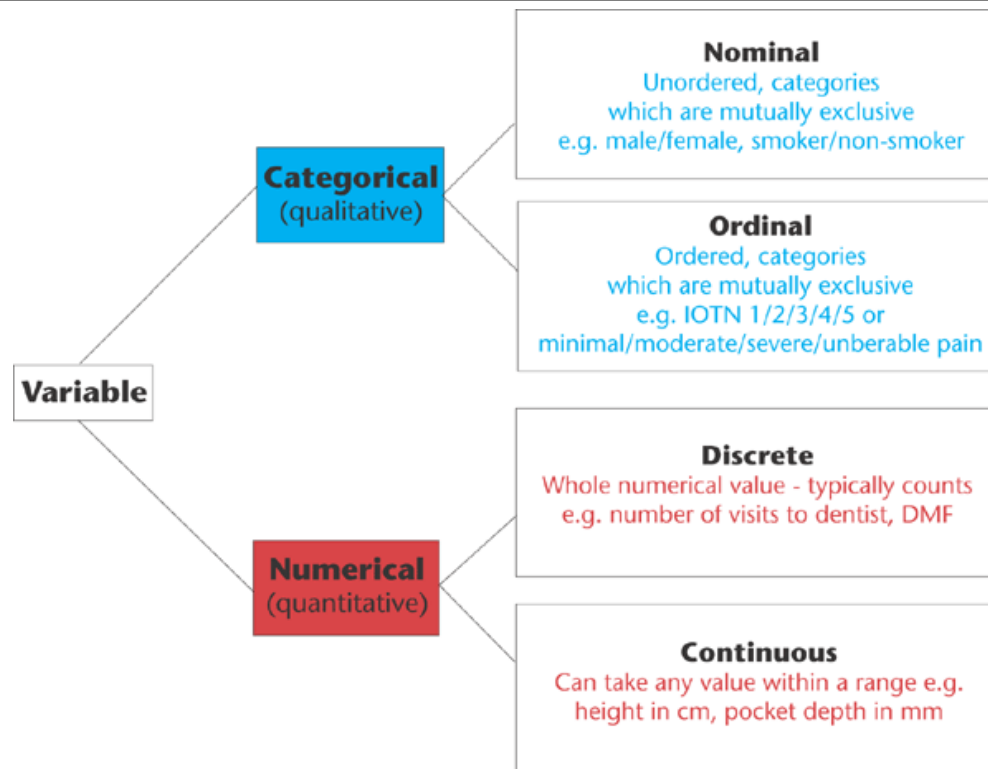## Artificial Creativity

ChatGPT: ChatGPT (Chat Generative Pre-trained Transformer) is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3 family of large language models, and is fine-tuned (an approach to transfer learning) with both supervised and reinforcement learning techniques.

o Question answer

o Solving math equations

o Writing texts (basic academic articles, literary texts, movie script, etc.)

o Interlingual translation

o Summarizing text and detecting keywords in text

o Classification

o Making recommendations

o Explaining what anything does (for example, explaining what a code block does)

Types of Data (Arial View)

Variable

Categorical (qualitative)

Nominal
Unordered, categories which are mutually exclusive
e.g. male/female, smoker/non-smoker

Ordinal
Ordered, categories which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or minimal/moderate/severe/unberable pain

Numerical (quantitative)

Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF

Continuous
Can take any value within a range e.g. height in cm, pocket depth in mm

Introduction to Python

What is Python?

Python is open source, general purpose and high level programming language

Applications?

Professionally being used for:
◦ Backend web development,
◦ Data analysis
◦ Artificial intelligence
◦ Scientific computing.

Also used for building:
◦ Productivity tools, games, and desktop apps

Introduction to Python

Why Python?
◦ Most popular introductory teaching language in US Universities
◦ It has been ranked **4th** in an IEEE survey for the most popular programming languages

Versions

Python 1.0 - January 1994

    Python 1.5, Python 1.6

Python 2.0 - October 16, 2000 - October 1, 2008

    Python 2.1 – Python 2.7
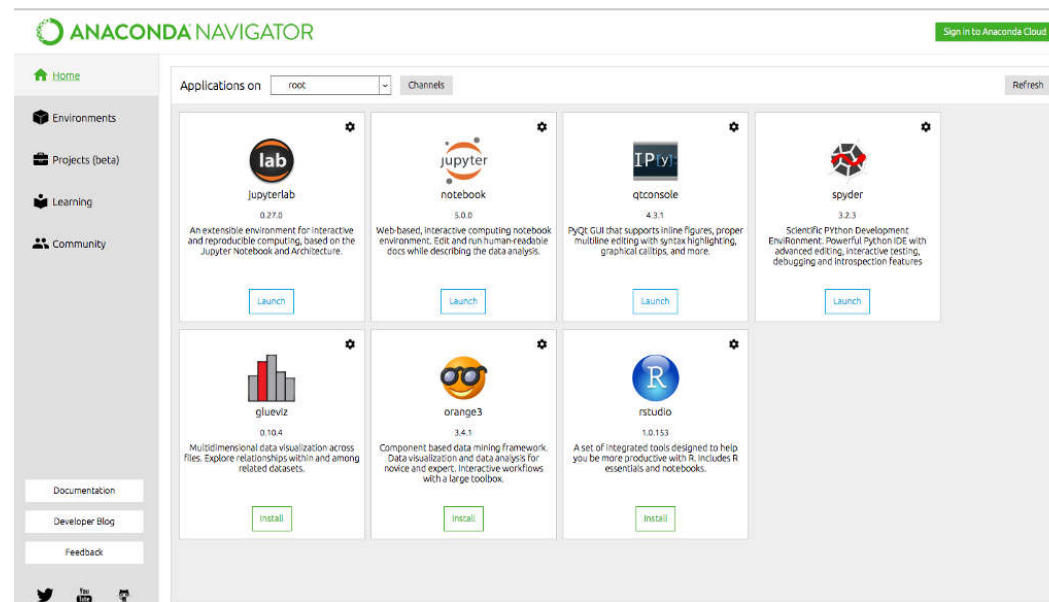
Python 3.0 - December 3, 2008 - till date

    Python 3.1 – Python 3.10.7

   ***** Current Version: **Python 3.10.7**

Anaconda

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system conda.

Anaconda - Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in **Anaconda**® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands.
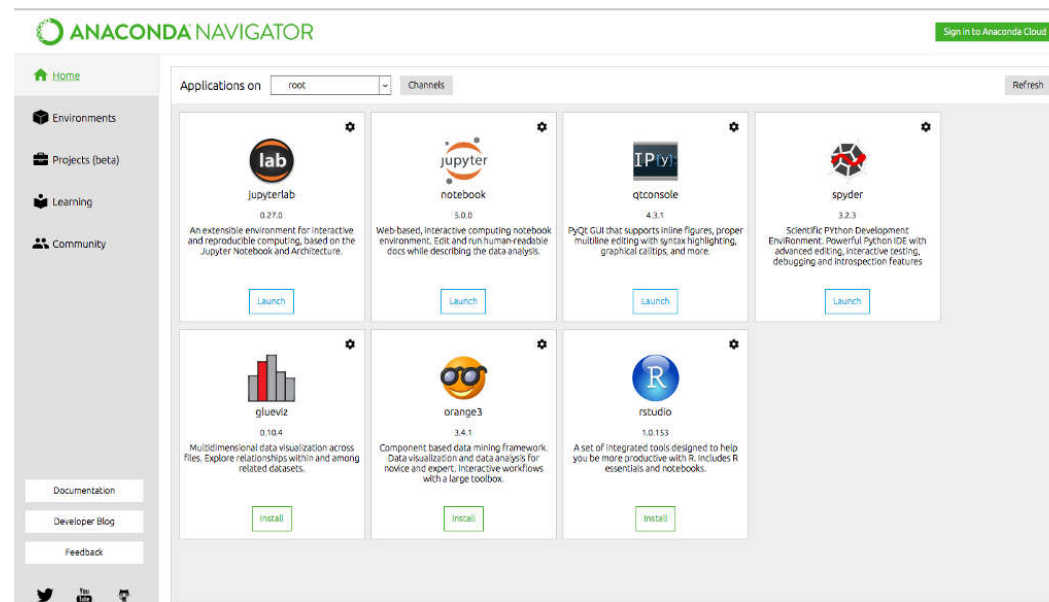
Anaconda – Jupyter Notebook

The **Jupyter Notebook** is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Anaconda – Spyder

Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates with a number of prominent packages in the scientific Python stack,including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, as well as other open source software. It is released under the MIT license.
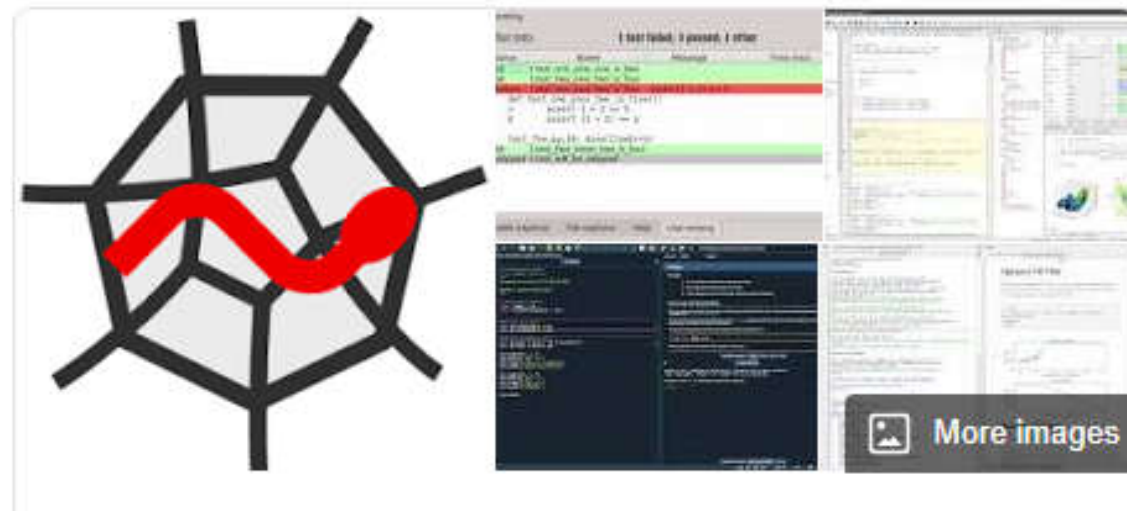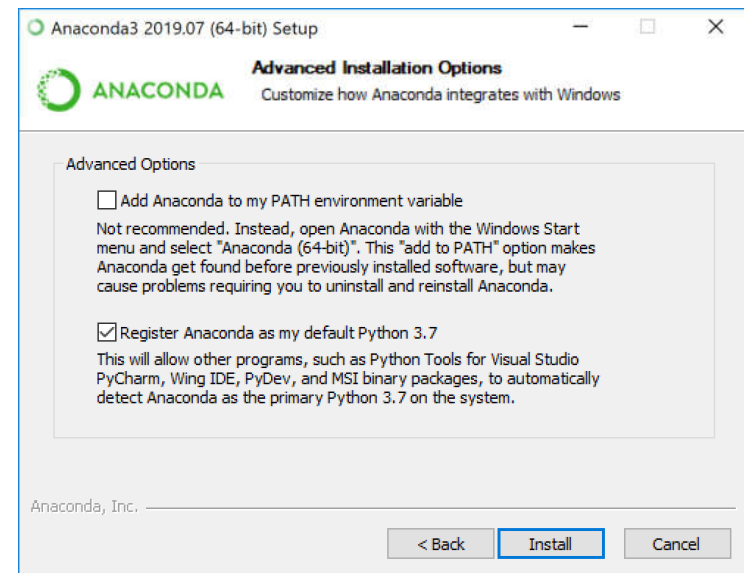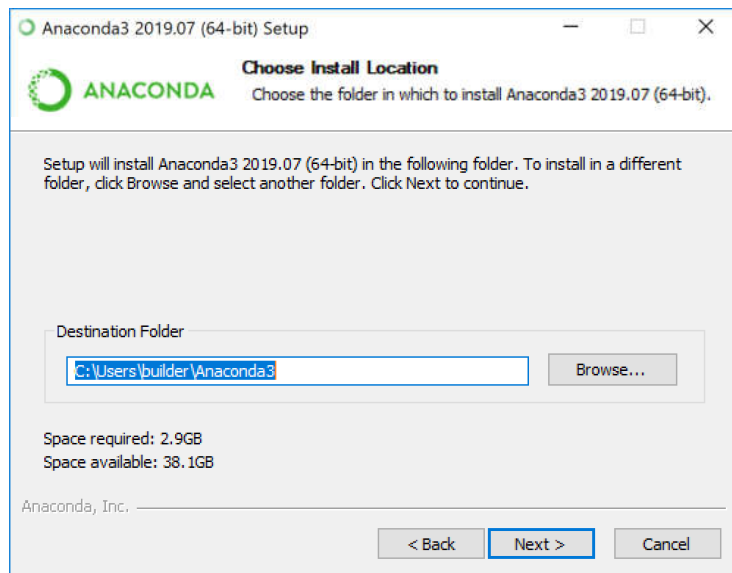
Anaconda - Installation

1- [Download the Anaconda installer](#)

2- Launch the installer and Select a destination folder to install Anaconda and click the Next button

Introducing Conda Environments

A conda environment is a directory that contains a specific collection of conda packages that you have installed.

For example, you may have one environment with NumPy 1.7 and its dependencies, and another environment with NumPy 1.6 for legacy testing. If you change one environment, your other environments are not affected.

You can easily activate or deactivate environments, which is how you switch between them.

You can also share your environment with someone by giving them a copy of your emvironment.yml file.

Creating Conda Environments

To create an environment with a specific version of Python:

```
conda create -n myenv python=3.7
```

To create an environment with a specific version of Python and multiple packages:

```
conda create -n myenv python=3.6 scipy=0.15.0
astroid babel
```

Creating an environment from an environment.yml file

Create the environment from the environment.yml file.

```
conda  env  create =f

environment.yml
```

EXAMPLE: A simple environment file:

```
name: stats
dependencies:
   - numpy
   - pandas
```

Export your active environment to a new file:

```
conda env export > environment.yml
```

Conda Basic Commands

Verify conda is installed, check version number	`conda info`
Update conda to the current version	`conda update conda`
Install a package included in Anaconda	`conda install PACKAGENAME`
Run a package after install, example Spyder*	`spyder`
Update any installed program	`conda update PACKAGENAME`
Command line help	`COMMANDNAME --help` `conda install --help`

Commands for Using Conda Environment

Create a new environment named py35, install Python 3.5	`conda create --name py35 python=3.5`
Activate the new environment to use it	`WINDOWS:     activate py35` `LINUX, macOS: source activate py35`
Get a list of all my environments, active environment is shown with *	`conda env list`
Make exact copy of an environment	`conda create --clone py35 --name py35-2`
List all packages and versions installed in active environment	`conda list`
List the history of each change to the current environment	`conda list --revisions`
Restore environment to a previous revision	`conda install --revision 2`
Save environment to a text file	`conda list --explicit > bio-env.txt`
Delete an environment and everything in it	`conda env remove --name bio-env`
Deactivate the current environment	`WINDOWS: deactivate` `macOS, LINUX: source deactivate`
Create environment from a text file	`conda env create --file bio-env.txt`

Commands for Installing/Updating Packages

Install a new package (Jupyter Notebook) in the active environment	`conda install jupyter`
Run an installed package (Jupyter Notebook)	`jupyter-notebook`
Install a new package (toolz) in a different environment (bio-env)	`conda install --name bio-env toolz`
Update a package in the current environment	`conda update scikit-learn`
Install a package (boltons) from a specific channel (conda-forge)	`conda install --channel conda-forge boltons`
Install a package directly from PyPI into the current active environment using pip	`pip install boltons`

pip VS conda Commands

Install and Manage Packages in Python

pip	conda ANACONDA
pip search pyserial	conda search pyserial
pip install pyserial	conda install pyserial
pip install pyserial --upgrade	conda update python
pip list	conda list

	conda	pip
install python package	✅	✅
create virtual environment	✅, built-in	❌, requires `virtualenv` or `venv`
package format	`.tar.bz2` , `.conda`	`.whl` , `.tar.gz`
manages	binaries	wheel or source
can require compilers	❌	✅
package types	any	Python-only
dependency checks	✅	❌
package sources	Anaconda repo and Anaconda cloud	PyPI

Python Basics: Outline

Module-I

Introduction, Syntax, Variables, Operators, Loops, Conditions, Arrays

Module-II

Functions, Default Value Arguments , Arrays, Exception Handling, File I/O

Module-III

Basic Statistics and List processing using Numpy