CSC461
INTRODUCTION TO DATA SCIENCE

Dr. Muhammad Sharjeel

https://muhmmadsharjeel.github.io/

01

BASICS OF DATA SCIENCE

WHO AM I?

- **Dr Muhammad Sharjeel**
  - Assistant Professor, Computer Sciences Department, CUI, Lahore Campus
  - PhD Computer Science – England
  - MS Computer Science – Australia
- 14 years of teaching experience, some of the courses I teach/have taught
  - Programming Fundamentals
  - Introduction to Data Science
  - Advanced Algorithm Analysis
  - Machine Learning
  - Introduction to Information and Communications Technologies
  - Introduction to Computing
  - Computing for Management
  - Introduction to Computer Programming
  - Data Structures and Algorithms
  - Wireless and Mobile Computing
  - Network Security
  - Design and Analysis of Algorithms
  - Data Security and Encryption
  - Information Security

MY TEACHING STYLE

- I like to
  - Interact with my students
  - Ask questions
  - Be asked questions
  - Give assignments and projects
- **I like to learn**
- I am very bad at memorizing names (not faces)

IMPORTANT INFORMATION

- Instructor: Dr Muhammad Sharjeel
- Email: muhammadsharjeel@cuilahore.edu.pk (Office – PhD Block - 186)

- Google Drive shared folder link https://tinyurl.com/fa23i2dc or scan the QR code

- Assessment:
  - Midterm = 25%
  - Quiz/Assignments = 25%
  - Terminal = 50%

COURSE AIMS

- To understand the basic concepts and principles of data science, making data-driven decisions and effectively communicating results
- To learn how to use programming tools for acquiring, cleaning, analyzing, exploring, and visualizing data
- To know how to collect data from unstructured sources and store it
- To analyze data rigorously using a variety of statistical and Machine Learning approaches
- To develop skills in designing and building professional data science applications

- A major component of this course is to learn how to use python programming to apply a variety of methods to real-life datasets

PRE-REQUISITES

- MTH262 - Statistics and Probability Theory
- Basic programming concepts
  - Python programming language

- This course is for you, if you want to
  - Apply data science in your own field
  - Work in industry or research
  - Understand data’s role in society

KEY TOPICS

- Basics of Data Science
- Data Collection and Scraping
- Relational Data, Visualization and Data Explorations
- Linear Algebra, Graph, and Network Processing
- Text Analysis and Natural Language Processing
- Statistical Modeling and Machine Learning
- Recommender Systems
- Big Data

WHAT IS DATA SCIENCE

- We live in a world that's drowning in data
- Data is now recognized as one of the founding pillars of our economy, and the notion that the world grows exponentially richer in data every day is already yesterday's news
- With so much driven by data, it's important that data scientists work responsibly and for the greater good

WHAT IS DATA SCIENCE

- “The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.” Hal Varian, Chief Economist at Google
- “Data science is the application of computational and statistical techniques to address or gain insight into some problem in the real world.” Zico Kolter, Professor at CMU

- The goal of data science is to extract insight and knowledge from large amounts of data or to find patterns within the data

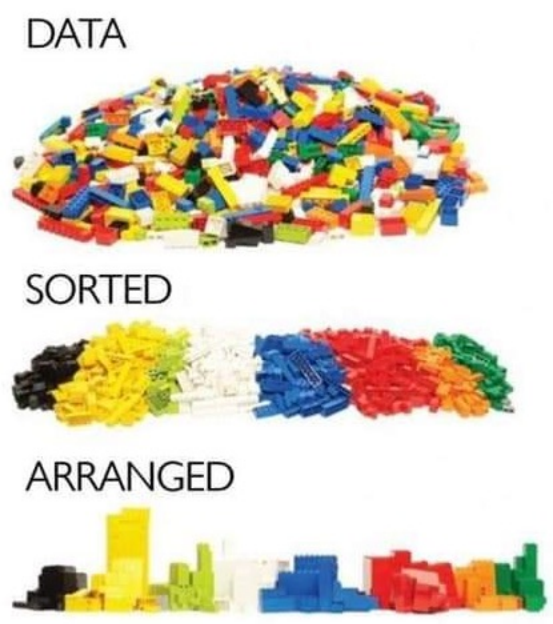
WHAT IS DATA SCIENCE

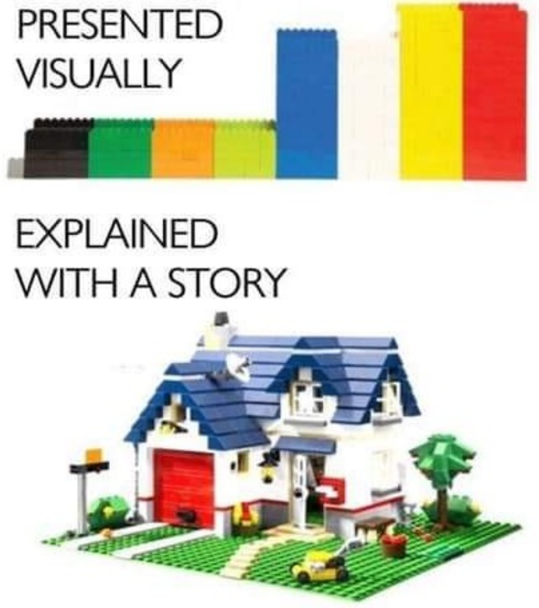
- Data science = statistics +
  data processing +
  machine learning +
  scientific inquiry +
  visualization +
  business analytics +
  big data

WHAT IS NOT DATA SCIENCE

- Data science is not Machine Learning
  - Machine Learning involves computation and statistics, but has not (traditionally) been very concerned about answering scientific questions
  - Machine Learning has a heavy focus on fancy algorithms, however, at times, the best way to solve a problem is just by visualizing the data
- Data science is not Statistics
  - Analyzing data computationally, to understand some phenomenon in the real world, that sounds an awful lot like statistics
  - Statistics has evolved a lot more along the mathematical/theoretical frontier
  - Not many statistics courses have a lecture on, e.g., web scraping, or a lot of data processing more generally
- Data science is not Big Data
  - Sometimes, in order to truly understand and answer your question, you need massive amounts of data, but sometimes you don’t
  - Don’t create more work for yourself than you need to

WHAT IS DATA SCIENCE

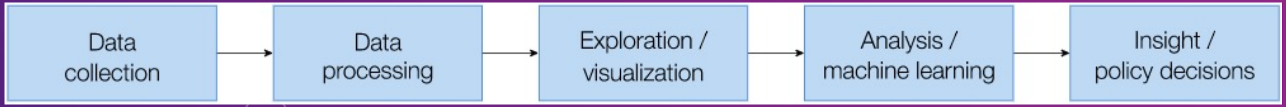


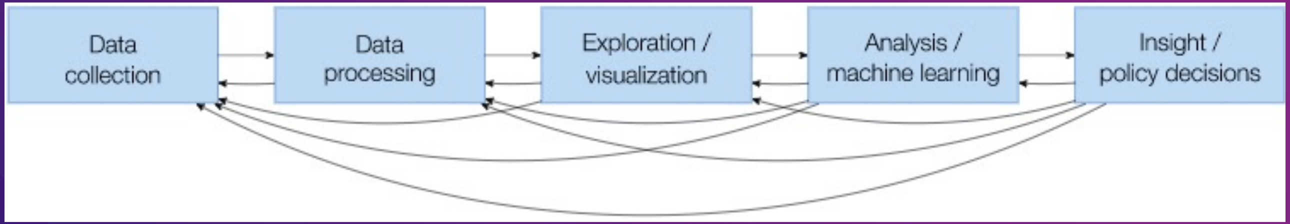


DATA SCIENCE APPLICATIONS

- Smart cities
- Healthcare
- Business analytics
- Biomedicine
- Social media networks
- Finance
- Natural sciences

DATA SCIENCE PIPELINE


Data collection
Data processing
Exploration / visualization
Analysis / machine learning
Insight / policy decisions


Data collection
Data processing
Exploration / visualization
Analysis / machine learning
Insight / policy decisions

PUBLICLY AVAILABLE DATASET REPOSITORIES

- https://www.data.gov/
- https://cloud.google.com/bigquery/public-data/
- https://www.kaggle.com/datasets
- https://aws.amazon.com/public-datasets/
- https://github.com/awesomedata/awesome-public-datasets

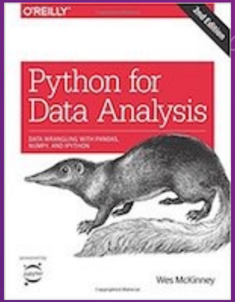
PYTHON - THE LANGUAGE OF DATA SCIENCE

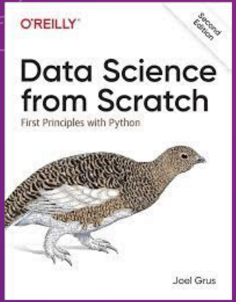
- Especially true, if the data science tasks involve lots of data processing and/or machine learning
- Less true, if the tasks are more “purely statistical” (then R is more standard)

- Gettings started with Python
  - https://jakevdp.github.io/PythonDataScienceHandbook/
  - https://github.com/donnemartin/data-science-ipython-notebooks
  - http://opentechschool.github.io/python-data-intro/

TEXTBOOKS

- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython
- Data Science from Scratch: First Principles with Python





WRAP-UP FOR TODAY

- Please get in touch with me if you’re unsure of whether or not you’re at the right level for this course
    - My guess is that you are!

- Read about Docker and Jupyter on the web
    - https://www.docker.com
    - https://www.jupyter.org
- Make a GitHub account
    - https://www.github.com/
- Download and install Anaconda (Python)
    - https://www.anaconda.com/download
    - https://www.python.org
- Start using Google Colab and get yourself familiarize with it
    - colab.research.google.com/

DATA SCIENCE – UP-TO-DATE

- Pakistan Data Science Hackathons
  - https://foundry.pk/data-competitions-hackathons/
- Data Science Pakistan Facebook Group
  - https://www.facebook.com/groups/datascipak/
- Harvard Data Science Review
  - https://hdsr.mitpress.mit.edu/

THANKS