

PRINCIPLES OF DATA MANGEMENT

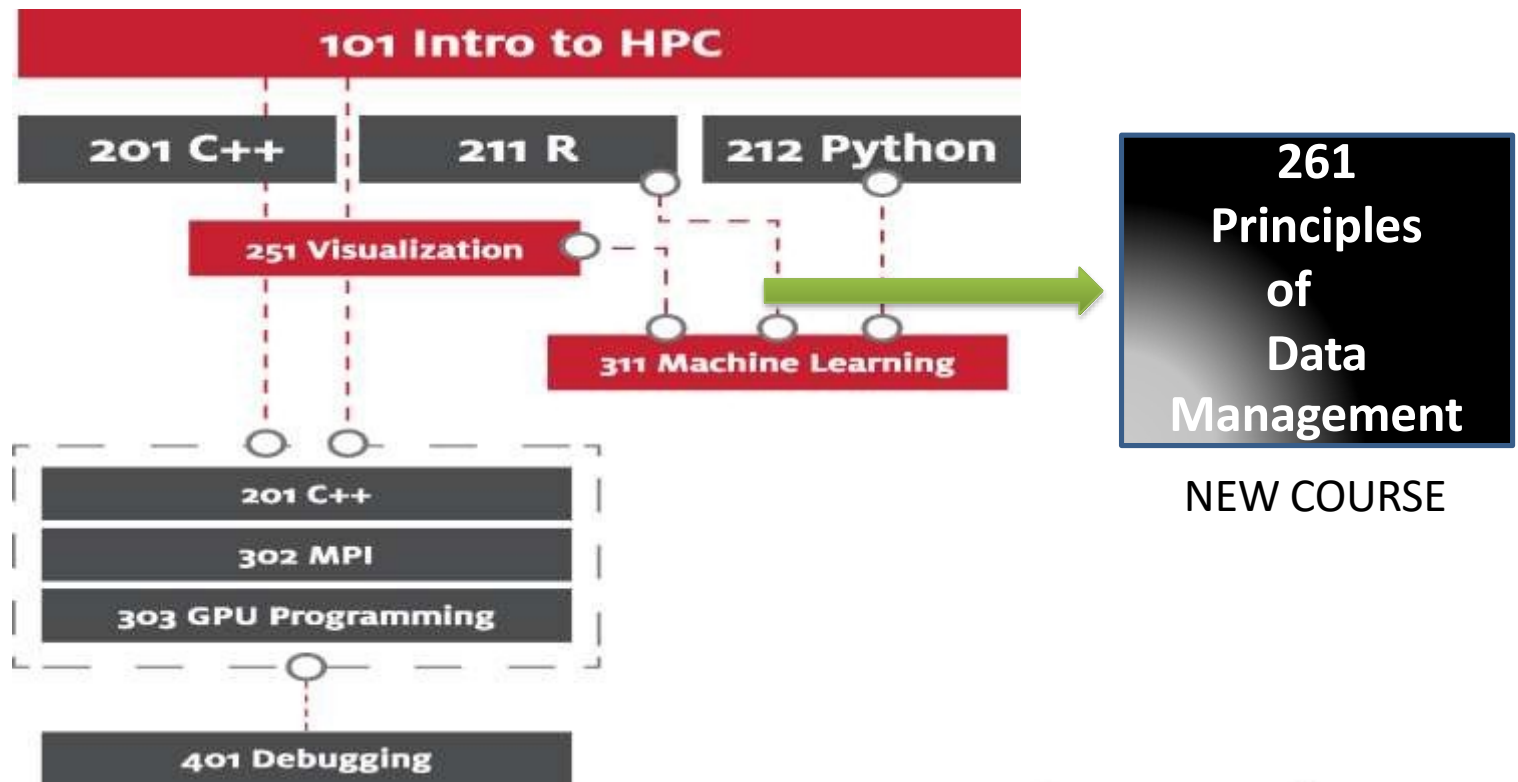
Instructor: Ishita Sharma

Email: isharma3@uh.edu

UNIVERSITY of
HOUSTON

DIVISION OF RESEARCH
HPE DATA SCIENCE INSTITUTE

HPE-DSI Training Courses



We also offer on-demand courses, e.g. face-to-face consultation
<https://hpedsi.uh.edu/education/training>

Course Overview



Business Problem



Data Extraction

Data Importing
Web Scraping



Data Management



Data Transformation

Data Preprocessing (Joining data tables, sorting data, missing values, time/date)



Exploratory Data Analysis



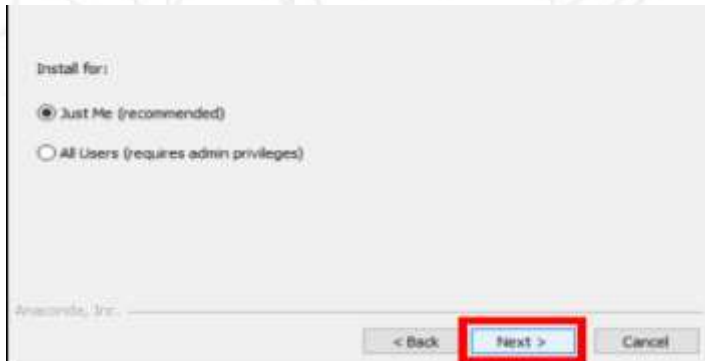
Statistics

Course Resources

- You will need a laptop/desktop
- This course is in Python
- The IDE(Integrated Development Environment) we are using for the course is Anaconda's jupyter notebook
- How to install Anaconda and start using your jupyter notebook?
 - [Anaconda Intsall Guide](#)
 - Choose your operating system(Mac/Windows/Linux)
 - Choose python version (Recommended 3.X)
 - Then locate the file where it has been downloaded in your computer
 - Right-click on the downloaded file >> Run administrator
 - Install by agreeing all the conditions



- There will two options



(Recommended – Just me)

- Check your install location and click Next
- Check options(Recommended – both check box) and Install
- After this Anaconda Navigator and Anaconda Prompt will be installed
- Open terminal in Mac or Command Prompt in Windows and to test type : jupyter notebook
- To locate the path:
 - where conda
 - where python



Problem RATHER Motivation

- COVID-19 Real time analysis
 - Air Traffic
 - Ecommerce stores
 - Uber use trends
 - Hotel affected
- Fraud Detection:
Insurance company dealing a major problem of insurance fraud
- Our Example: Austin Airbnb Dataset
 - We will be looking at the different Airbnb available in different year.
 - Different prices of the room types based on the zip code.

What do you understand by "Data Management"?

➤ What:

- Data management is the practice of managing data as a valuable resource to unlock its potential for an organization. Managing data effectively requires having a data strategy and reliable methods to access, integrate, cleanse, govern, store and prepare data for analytics – **SAS**
- Data management refers to the professional practice of constructing and maintaining a framework for ingesting, storing, mining, and archiving data integral to modern business – **talend**.

➤ Why:

- Data Quality
- Data Governance
- Data Analysis
- Data Presentation

Git & Git-Hub

➤ Usage

- to keep the projects organized
- Collaborate with the team project

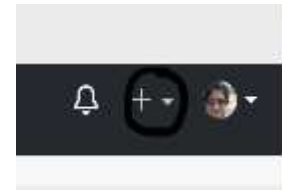
➤ Installing Git & Creating a GitHub Repository

- To install Git : [git-scm download](#)

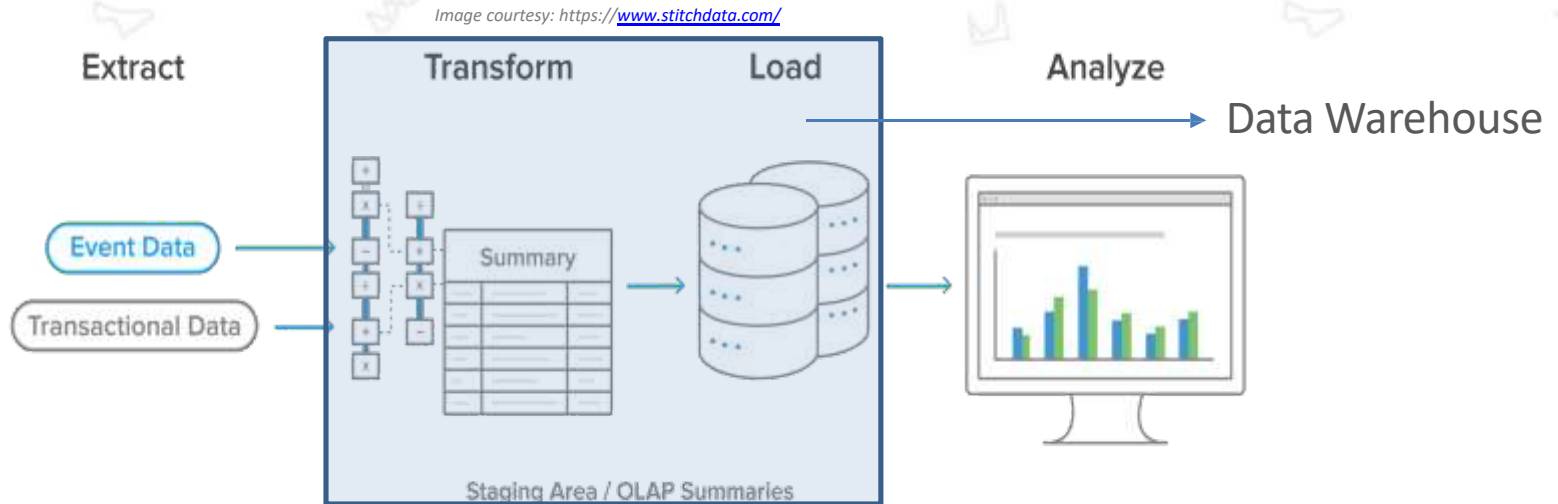
For Windows/Linux/ Mac

- **Create GitHub Repository:**

- Go to GitHub website and sign up or sign in(already account)
<https://github.com/>
- Next step clicks on the '+' on the top-right of your profile picture
- Git Commands on the terminal(for mac)
 - git config --global user.email "[yourGitHub@email.com](#)"
 - git config --global user.name "yourGitHubusername"

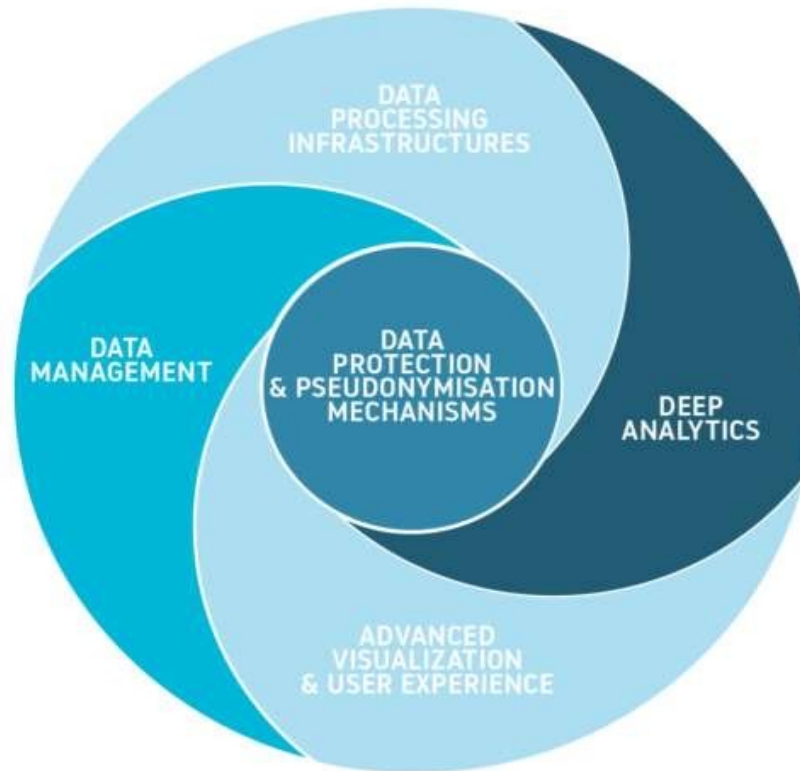


Extract Transform Load(ETL/ELT)



- Extract: This is first step where someone in the organization identifies the desired data sources based on the **business problem** and the rows, columns, and fields to be extracted from those sources. These sources likely include:
 1. Transactional databases hosted on-site or in the cloud
 2. Hosted business applications
 3. API and sensors
- Transform: The second important step is data treating and data cleaning.
- Load: This step the data is ready to be loaded in a data warehouse
- Analyze: Last step data is further analyzed using the BI tools such as tableau/plotly

Data Scientist: life cycle



Data

A Data Dictionary provides metadata about data elements

The metadata assist in defining the scope, rules for their usage and application of the data

Important data documents:

➤ Meta Data:

- Metadata is data about data. Information describes something like
 - web page
 - document
 - license
 - file size
 - author

➤ Data Dictionary:

- It is a collection of
 - column names
 - definitions
 - attributes
- It describes the meanings and purposes of data elements within the context of a project, and provides guidance on interpretation, accepted meanings and representation

DATA SCIENTIST TOOLBOX

We are using **Python** for this course.

Why Python?

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages.

- Data Evaluation
- Data Extraction and Transform
 - Pandas
 - NumPy
- Data Exploration
 - Matplotlib
 - Seaborn



What does a Data Scientist desire?

- **Data**
- **Tabular Heterogeneous Data Structure**
- **Data Treatment:**
 - Data Reviewing
 - Data Cleaning
 - Data transforming
 - Data Aggregating
 - Data Slicing and indexing
- **Data Exploring**
- **Applying Statistics**
- **Data Visualization**
- **Creating Models**
- **Feature Engineering**



Machine Learning

Pandas:

```
import pandas as pd(alias name)
```

- Pandas built on top of NumPy
- Pandas and NumPy belong to SciPy
- Pandas comes to rescue our analysis by providing various features such as:
 - Provides data structure to work

Series

Data Frame(table-like structures as in R)

- Provides tools for data treatment
- Very efficient with data upto 1GB

Limitation: Performance may be concern for very large data.

- https://pandas.pydata.org/docs/user_guide/index.html

NumPy:
import numpy as np(alias name)

- NumPy most important feature memory usage.
- NumPy uses BLAS(Basic Linear Subroutines) in its backend
- Important:
 - Provides faster vectorized operations on multidimensional arrays and matrices such as transpose , dot products etc.
 - Provides numeric data filtering
- <https://numpy.org/>

Data Extraction: Part 1



➤ Flat Files:

- Store's data in a plain text file.
- Each line of the file holds one record
- The fields are separated by delimiters, such as commas or tabs.
- `pd.read_csv('file_name.csv')`
(If your files are large, you can use chunksize parameter)
- `pd.read_table('filename/URL')`

➤ Other Files:

- Excel files: `pd.read_excel('file.xlsx', sheet_name = 'Sheet1', index_col = None, na_values = ['NA'])`
- Json file format: `pd.read_json('file.json')`
- SaaS file format: `pd.read_sas('file.sas7bdat')`

<https://pandas.pydata.org/pandas-docs/stable/reference/io.html>

Data Extraction: Part 2

➤ API(Application Programming Language)

Wikipedia terms that: “ In [computer programming](#), an API is a set of [subroutine](#) definitions, [protocols](#), and tools for building [application software](#).

Generally, it is like a communication between your request, and it sends the request to website and if approved >> desired data is extracted

- Output of REST API
 - *Json*
 - *Xml*

Example URL:

<https://api.exchangeratesapi.io/latest?symbols=USD,INR>

Query
Question
Mark

values

XML	JSON
<pre><Node> <id>10002</id> <Name>John</Name> </Node> <Node> <id>10003</id> <Name>Scott</Name> </Node> <Node> <id>10004</id> <Name>Mohan</Name> </Node> <Node> <id>10001</id> <Name>Deepak</Name> </Node></pre>	<pre>[{ "id":10002, "name":"John" }, { "id":10003, "name":"Scott" }, { "id":10004, "name":"Mohan" }, { "id":10001, "name":"Deepak" }]</pre>

Web Scrapping

➤ What:

- Process of extracting required data from the webpage
- Data is then processed in a data pipeline and stored in a structured format

➤ Where:

- Useful for Real Estate businesses to get the data of new projects, resale properties, available properties
- Hotel data
- E-commerce data to check the sale of different products

➤ Is it legal to scrape or not: [Legal Rules](#)


➤ Python Libraries install:

1. pip install *beautifulsoup4*
2. pip install *requests*
3. pip install *lxml*

When legal

1. Computer Fraud and Abuse Act (CFAA)

- When you don't access the data in an abusive fashion, you are legally safe. Secondly, as long as you don't use the same data for commercial purposes, you are not in violation of the CFAA.



- CFAA doesn't have abusive access and use of web data, violates the law, particularly for business purposes or financial gain. So when you indulge in web scraping in a way that violates the CFAA, your web scraping activity can be deemed illegal.

2. Copyright Infringement

- If you scrape the data but don't use the same data for adding on the internet or re-using it for commercial purposes, you are safe. Scraping is not illegal but re-using the copyrighted data for business purposes would be considered a violation of copyright laws.



- Companies may have data protected by copyright. Processing and using this data for commercial purposes may invite legal trouble.

3. Trespass to Chattel

- As long as you don't enter the prohibited space and don't behave in a way that harms the website in any way, it would be mostly legal.



- It's like trespassing on somebody else's property. In this case, you are entering a prohibited digital space and behaving in an irresponsible and harmful way.
- Entering the prohibited space on a website and behaving in a way that harms the website can turn a legal offense.

4. Robots.txt


- As long as you follow or respect the rules of Robots.txt, it is legal.
- If Robots.txt clearly prevents you from crawling or scraping, you need to ask for permission to access from the owner of the site before you go ahead and scrape the data.



- When you crawl and scrape and violate the norms laid down in Robots.txt, it becomes illegal. Not following or respecting Robots.txt can invite legal trouble.

5. Crawl Rate


- When you use a reasonable crawl rate and don't harass the site with repeated requests, your web scraping would be deemed legal.
- You should also follow the crawl-delay settings provided in Robots.txt.
- If there's no specific setting or rate mentioned, you should follow a conservative crawl rate of 1 request per 10-15 seconds.



- Websites are made for human use. So they are designed to handle a reasonable crawl rate. If you get aggressive and start using a fast crawl rate that brings down the server or harms the website, your actions become illegal.

6. API vs. scraping the data

- When you use an API if one is provided, instead of scraping data, it is perfectly legal.



- When you don't use the API given and behave in a way that harms either copyright laws or harm the website in any way, it becomes illegal.

7. Violating Terms of Service (ToS)


- When you follow and respect the Terms of Service (ToS), you are legally safe.
- If ToS clearly mentions that web scraping is not allowed, you should seek permission to access and scrape only after you get the permission.



- When you go against the Terms of Service (ToS) and scrape web data in a way that violates its terms and harms the website or the business in any way, it becomes illegal.

8. Hitting the servers too frequently


- When you access the website with reasonable a rate interval in between, it would be considered legal.
- You should also ensure that you don't send too many parallel requests and keep the number of parallel requests in control.



- When you don't follow a reasonable crawl rate and hit the server with frequent requests, leading to some sort of harm in the server or website, it can result in a legal problem.

9. Going beyond the Public Content

- As long as you scrape the data given in the public domain, you would be safe.
- If you don't scrape or re-publish it for financial gain, it would be legal.



- When you go beyond the public content and attempt to scrape the prohibited data and scrape it for business or financial gain.

When legal to scrape and when not?
Disclaimer: **Do not hack a website**
Follow Rules

Image courtesy: <https://prowebscraper.com/blog/is-web-scraping-legal/>

Data Frame(2D)

➤ Tabular Data Structure :

`pandas.DataFrame(data=None, index=None, columns=None, dtype=None, copy=False)`

Two-dimensional, size-mutable, potentially heterogeneous tabular data.

(<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>)

Series
created using
`np.array()`

age
39
50
38
53
28

Series

	age	work-class	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female

DataFrame

Data Files Export

➤ Excel

- `df.to_excel("output.xlsx",sheet_name='Sheet_name_1')`
- Documentation: [to_excel](#)

➤ Json

- `df.to_json(orient="split")`
- Documentation: [to_json](#)

Data List

Public Data sites for practice:

- [UCI MACHINE LEARNING REPOSITORY](#)
 - [Wine Dataset](#)
 - [Iris Dataset](#)
- [kaggle datasets](#)
 - [fatal-police-shootings-data.csv](#)
- <https://data.world/adamhelsinger/unicef-drinking-water-database>
- [data.gov](#)
 - [healthcare-cost-and-utilization-project-hcup-national-inpatient-sample](#)

Exercise

- Please select one of the data either from the links or by your own choice.
ONLY PUBLIC DATA PLEASE
- Please email me the data-set chosen - isharma3@uh.edu
(I can check for if the data-set is publicly available and if it is not too difficult to use)
- After selecting data:
 - Read the Meta-data and columns description
 - Think of two-three questions/problems which you can take out from data