Course Title:

Data Science (数据科学)

(Semester: Fall 2021)

Dr. Oluwarotimi W. SAMUEL

Research Center for Neural Engineering
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences

Contact: (Email: samuel@siat.ac.cn & timitex92@gmail.com)

Phone: +86-15814491870

(2021.10.09)

Data Sources, Collection and Formatting

❑Outline for today’s lecture

✓Data Sources for DS Projects

✓Data Collection Modalities

✓Data Formatting Tips

Data Sources, Collection and Formatting

□ **Objective:** This lecture will focus on discussing data sources, data gathering methods, and highlight some approach to data formatting.

□ **Expectation:** At the end of this lecture, students are expected to understand various data sources, gathering methods, and basic data formatting steps in the context of DS project

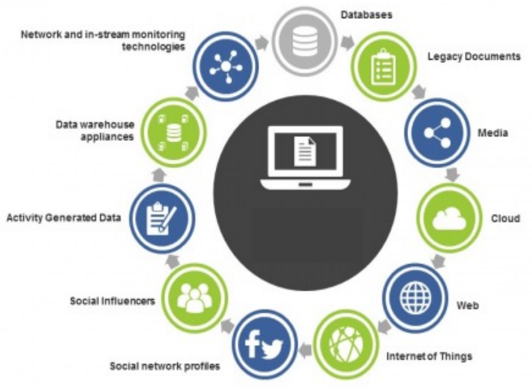
Data Sources, Collection, and Formatting

- ❑In developing a Data Science driven solution, you need to understand the problem **(First step)**.

Understanding the Problem

- ❑Next, you need to ask the following question: **What kind of data do I need to develop the solution?**

You need to decide the kind of data required for the project...



Data Sources

Data Sources

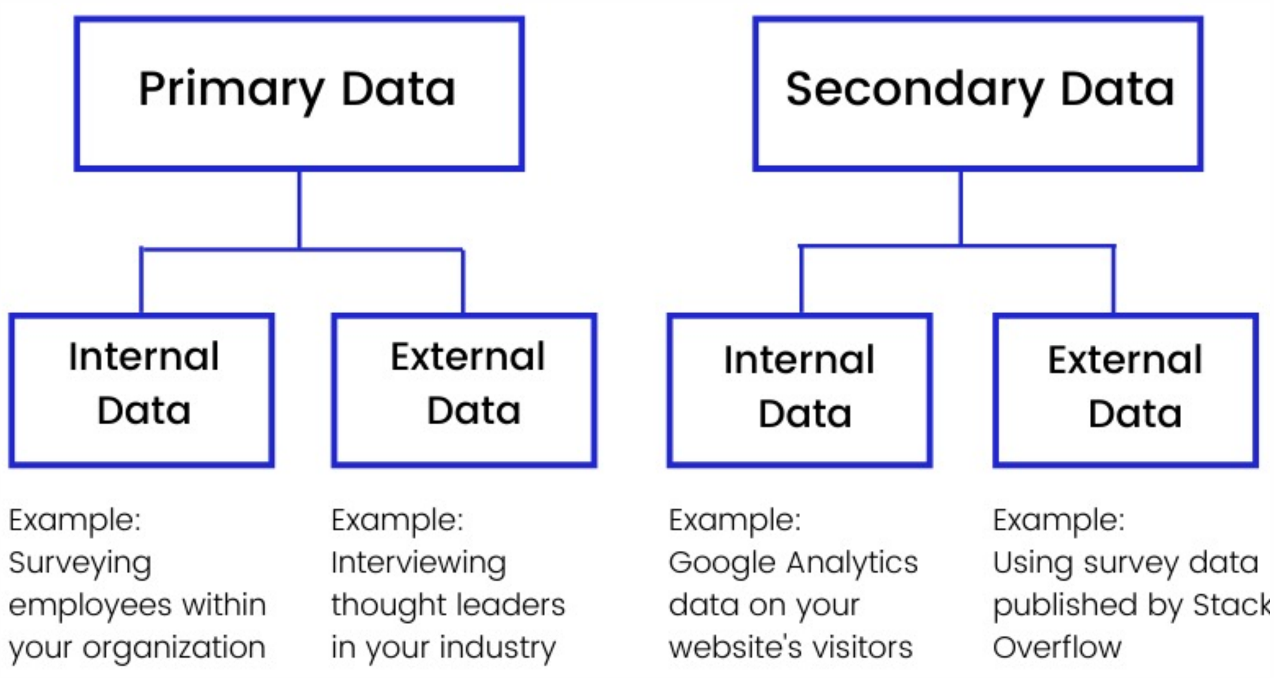
❑ How can I get the needed data for my project? This leads to asking about the Source of Data required.

- ✓ Primary Data Source
- ✓ Secondary Data Source

Data Sources

❑The flow diagram below shows a representation of the two categories of data sources and their respectively description.

Data Sources



Primary Data

Internal Data

Example: Surveying employees within your organization

External Data

Example: Interviewing thought leaders in your industry

Secondary Data

Internal Data

Example: Google Analytics data on your website's visitors

External Data

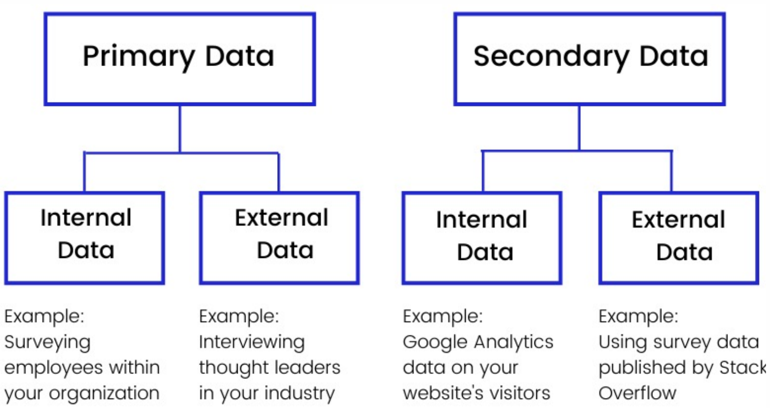
Example: Using survey data published by Stack Overflow

Data Sources

❑ Secondary Data Sources can be described as either Open Data Sources or Online Data Sources.

✓ Open Data (Always open for access, **does not require any subscription**)

✓ Online Data (Maybe accessed based on subscription or it may be freely accessed)


Primary Data
Secondary Data
Internal Data
External Data
Internal Data
External Data
Example: Surveying employees within your organization
Example: Interviewing thought leaders in your industry
Example: Google Analytics data on your website's visitors
Example: Using survey data published by Stack Overflow

Data Gathering/Collection

- ❑ Upon proper identification of the type of Data needed and its source (s), next is to design an approach via which such data can be collected.

- ❑ Data collection or gathering for DS project can be achieved via various means particularly with respect to the ***type of data*** and ***its source***.

- ❑Typically, data can be gathered in one of the following ways:

Data Gathering Methods

1 Direct download of data file (or files) manually

2 Query data from a database

3 Query an API (usually web-based)

4 Scrap data from a webpage

5 Acquisition of data by oneself

Data Gathering Methods

1 Direct download of a data file (or files) manually

This involves direct download of data in the form of data files from:

✓ "***within***"

or

✓ "***outside***"

an organization of interest. That is, such data can be downloaded mainly from 2 sources:

Data Gathering Methods

1 Direct download of a data file (or files) manually

That is, such data maybe obtained from:

- ✓ Internal Repository (private): Mostly owned and managed by a specific company.

- ✓ External repository (publicly available): Can be freely accessed

Data Gathering Methods

1 Direct download of a data file (or files) manually

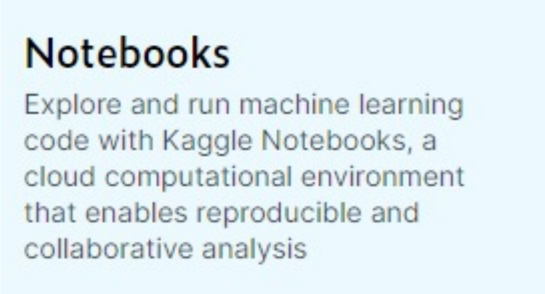
Private Repository

Data Gathering Methods

1 Direct download of a data file (or files) manually

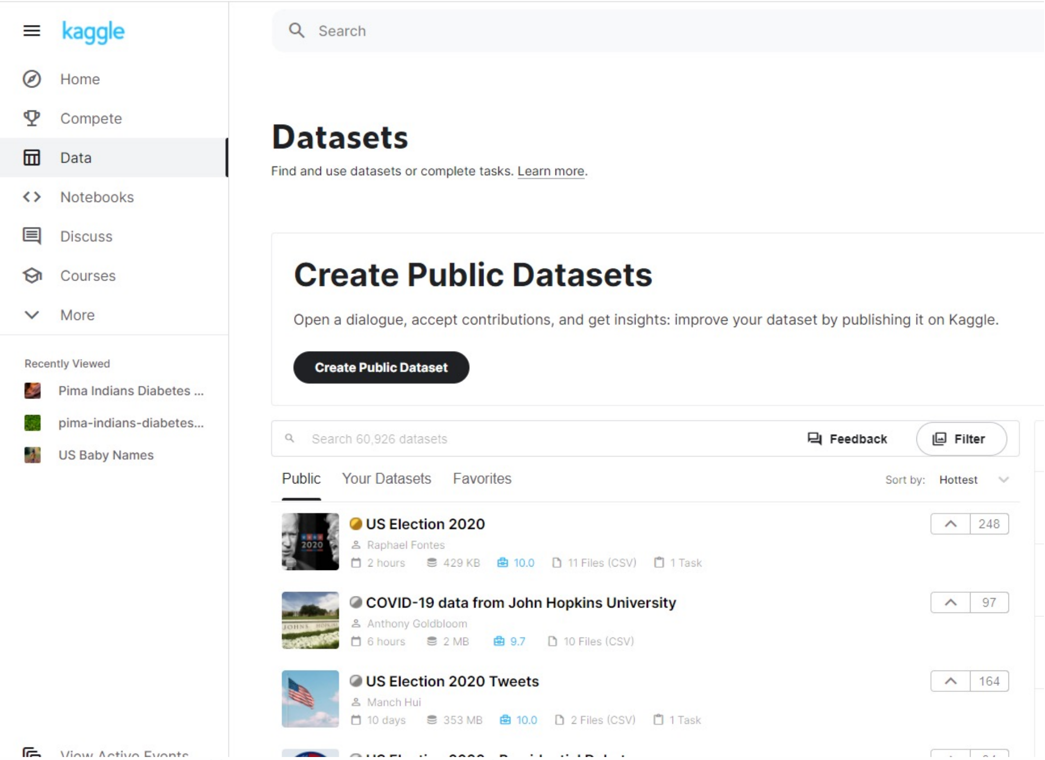
Public Repository

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.



Notebooks

Explore and run machine learning code with Kaggle Notebooks, a cloud computational environment that enables reproducible and collaborative analysis



Datasets

Find and use datasets or complete tasks. Learn more.

Create Public Datasets

Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle.

US Election 2020 — Raphael Fontes — 2 hours — 429 KB — 10.0 — 11 Files (CSV) — 1 Task — 248

COVID-19 data from John Hopkins University — Anthony Goldbloom — 6 hours — 2 MB — 9.7 — 10 Files (CSV) — 97

US Election 2020 Tweets — Manch Hui — 10 days — 353 MB — 10.0 — 2 Files (CSV) — 1 Task — 164

https://www.kaggle.com/datasets

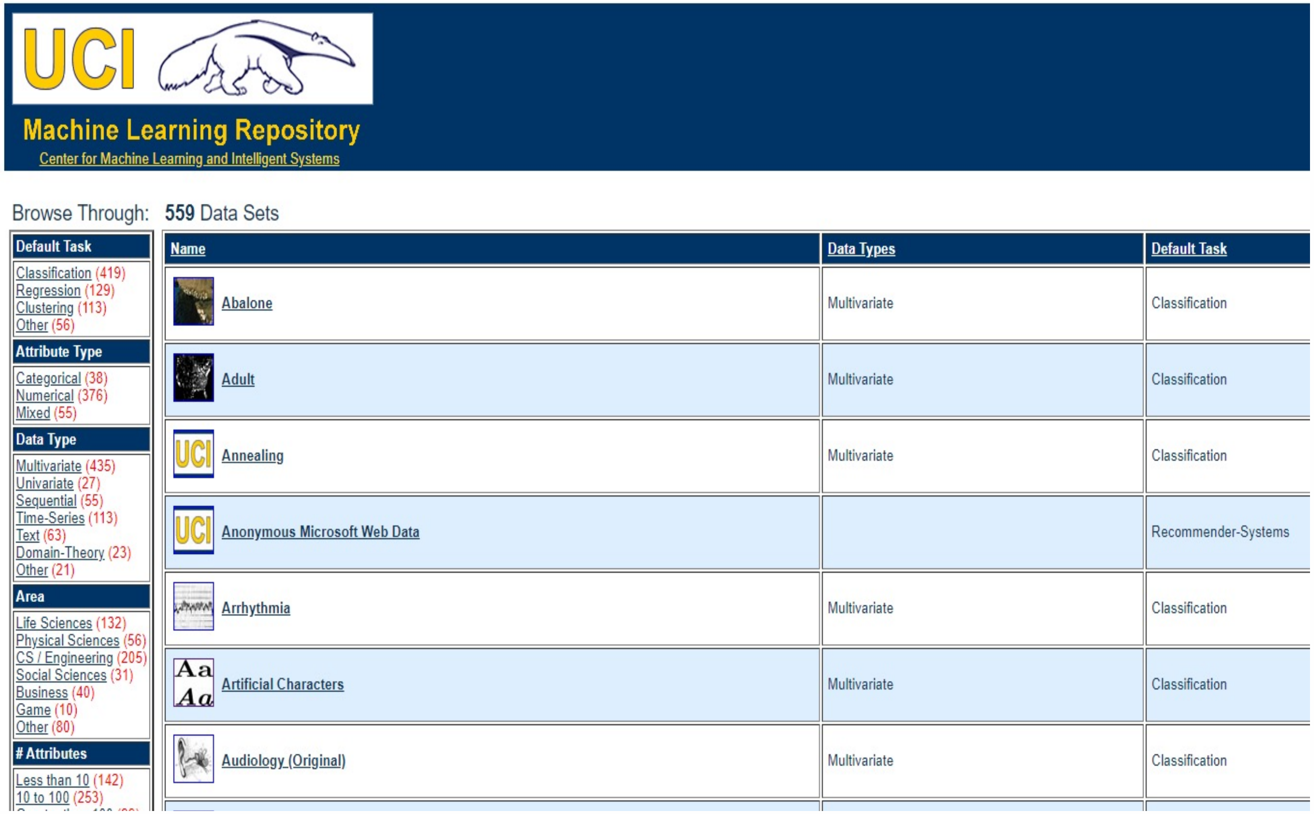
https://www.kaggle.com/kumargh/pimaindiansdiabetescsv

Data Gathering Methods

1 Direct download of a data file (or files) manually

Public Repository

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine.



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Browse Through: 559 Data Sets

Default Task
Classification (419)
Regression (129)
Clustering (113)
Other (56)

Attribute Type
Categorical (38)
Numerical (376)
Mixed (55)

Data Type
Multivariate (435)
Univariate (27)
Sequential (55)
Time-Series (113)
Text (63)
Domain-Theory (23)
Other (21)

Area
Life Sciences (132)
Physical Sciences (56)
CS / Engineering (205)
Social Sciences (31)
Business (40)
Game (10)
Other (80)

# Attributes
Less than 10 (142)
10 to 100 (253)

Name	Data Types	Default Task
Abalone	Multivariate	Classification
Adult	Multivariate	Classification
Annealing	Multivariate	Classification
Anonymous Microsoft Web Data		Recommender-Systems
Arrhythmia	Multivariate	Classification
Artificial Characters	Multivariate	Classification
Audiology (Original)	Multivariate	Classification

https://archive.ics.uci.edu/ml/datasets.php

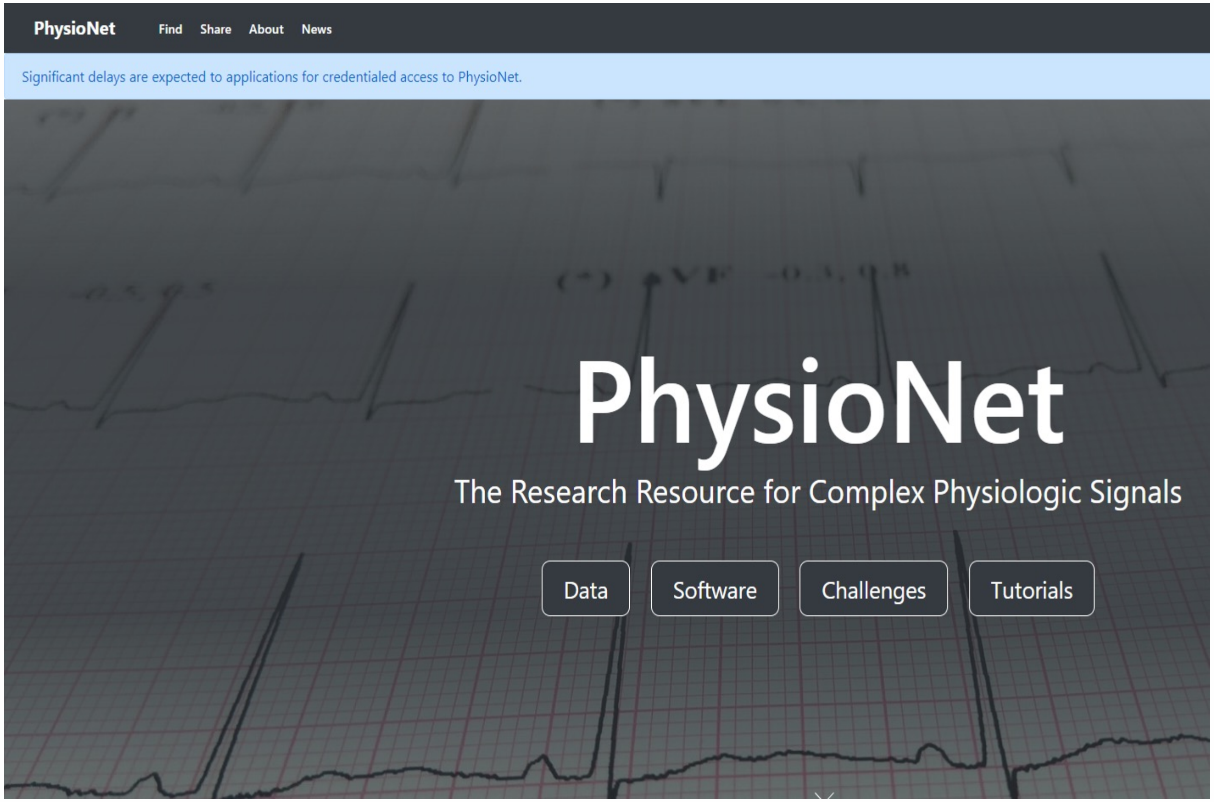
Data Gathering Methods

1 Directly download a data file (or files) manually

Public Repository

PhysioNet, the moniker of the *Research Resource for Complex Physiologic Signals*, was established in 1999 under the auspices of the National Institutes of Health (NIH), as described further below. The PhysioNet Resource's original and ongoing missions were to conduct and catalyze for biomedical research and education, in part by offering free access to large collections of physiological and clinical data and related open-source software.

https://physionet.org/#latest



Significant delays are expected to applications for credentialed access to PhysioNet.

PhysioNet

The Research Resource for Complex Physiologic Signals

Data Gathering Methods

2 Query data from a database

This involves writing simple/complex queries with the aid of a programming language to access data from a database of interest.

A typical database query language is the SQL (Structured Query Language) that allows access to the data.

Data Gathering Methods

2 Query data from a database

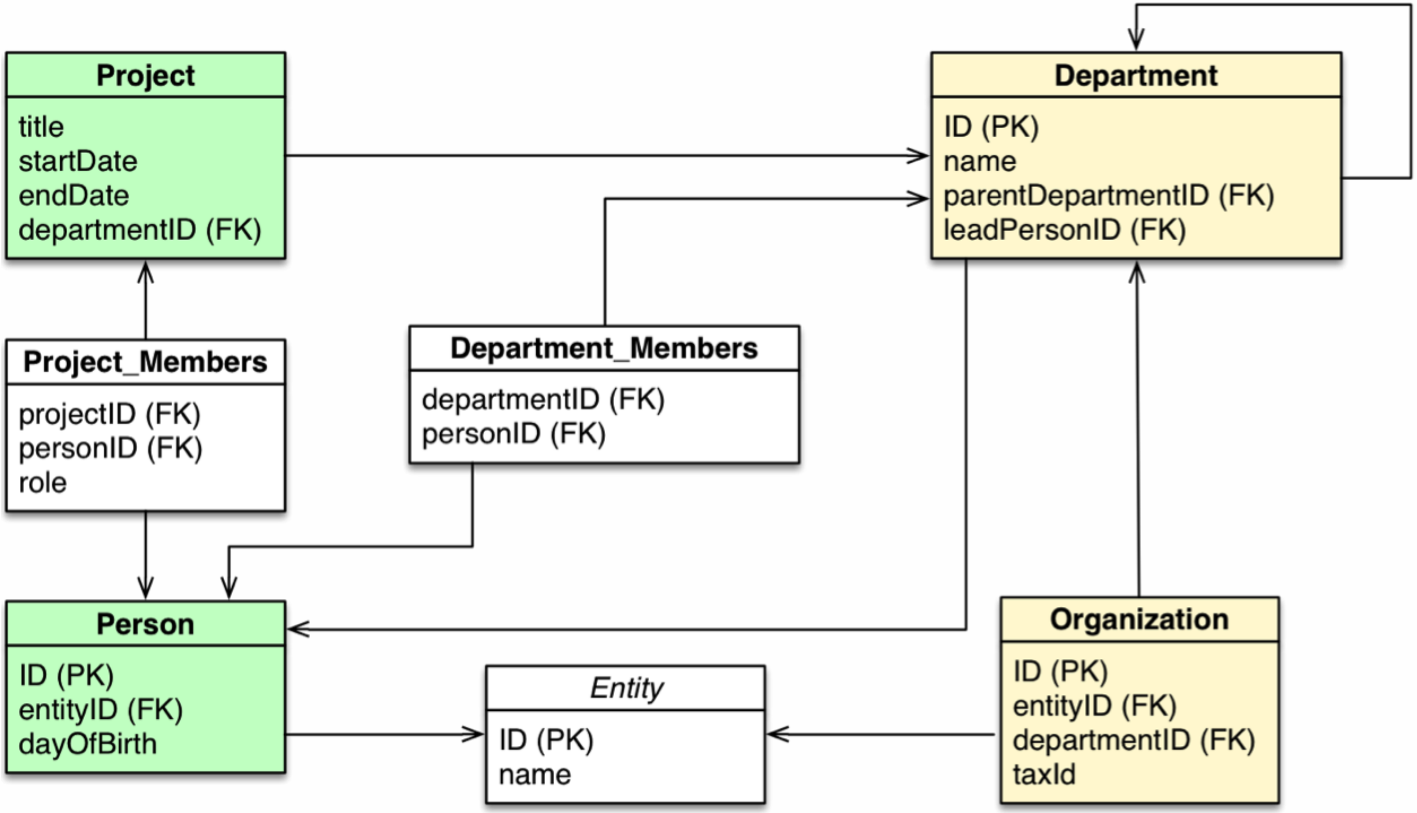
It is important to know that queries (SQL commands) are meant to operate on structured data, mostly organized in tabular forms.

The tables that are contain data and are related in a way that records can be easily retrieved or stored or modified.

Data Gathering Methods

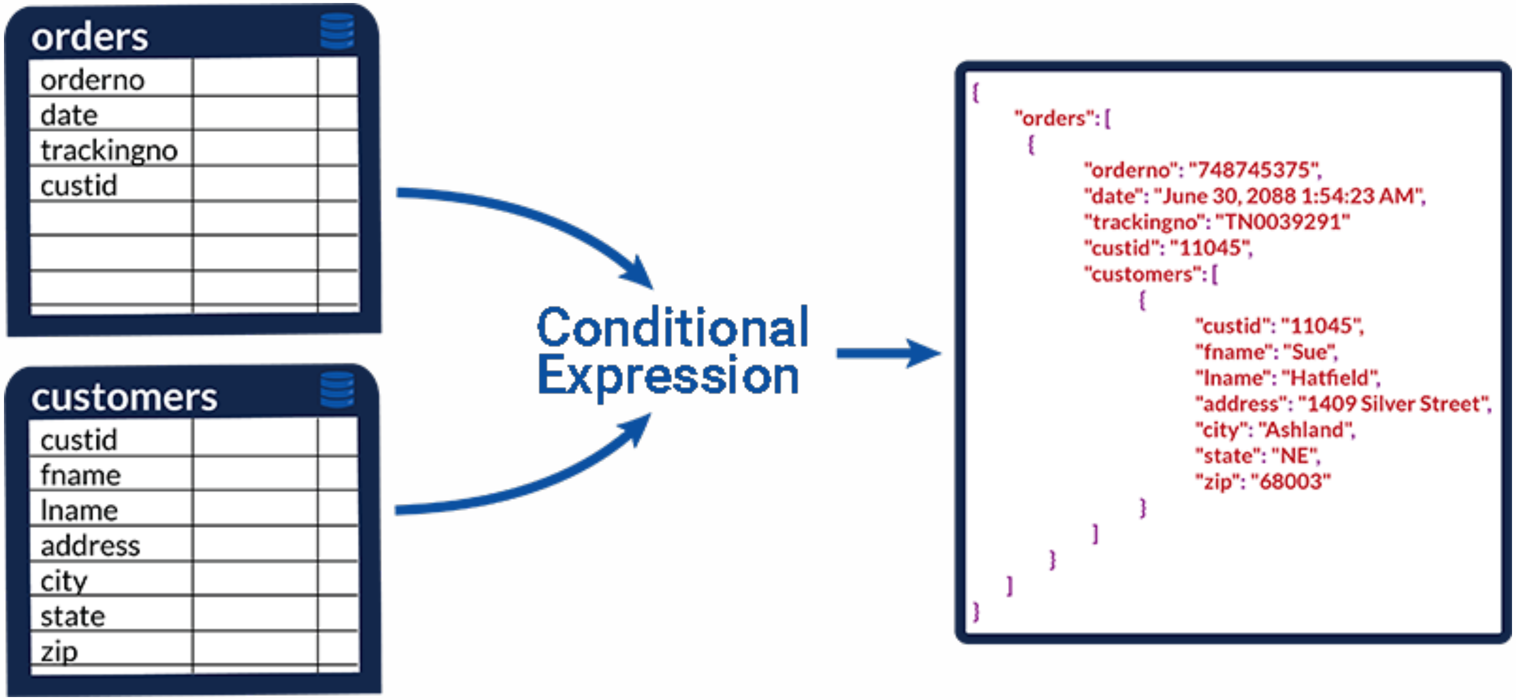
2 Query data from a database

Database Structure



Data Gathering Methods

2 Query data from a database



orders

orderno		
date		
trackingno		
custid		

customers

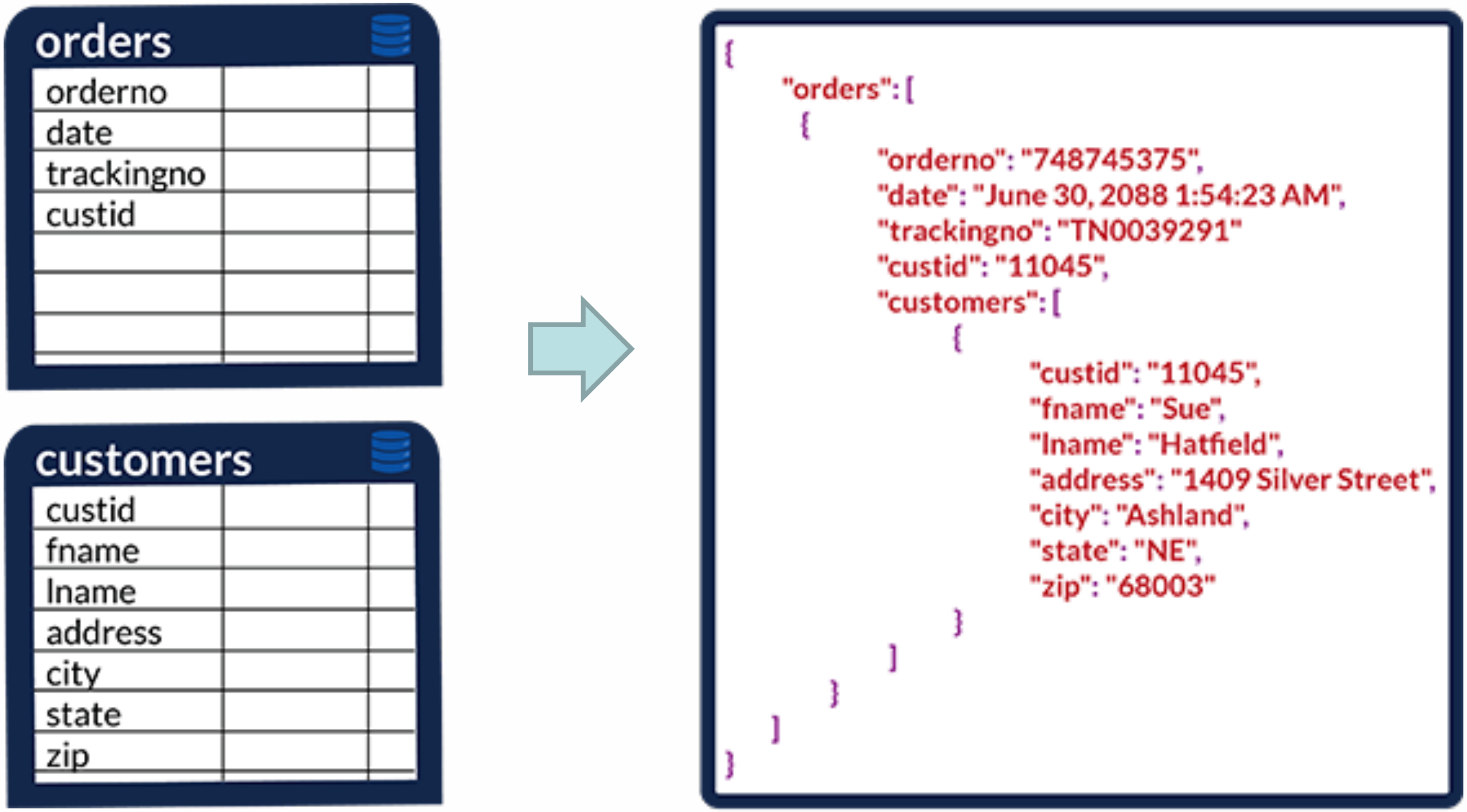
custid		
fname		
lname		
address		
city		
state		
zip		

```
{
    "orders":[
      {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 AM",
            "trackingno": "TN0039291"
            "custid": "11045",
            "customers":[
                  {
                        "custid": "11045",
                        "fname": "Sue",
                        "lname": "Hatfield",
                        "address": "1409 Silver Street",
                        "city": "Ashland",
                        "state": "NE",
                        "zip": "68003"
                  }
              ]
         }
     ]
}
```

https://www.goanywhere.com/managed-file-transfer/more/tutorials/how-to-query-a-database-and-write-the-data-to-json

Data Gathering Methods

2 Query data from a database



orders		
orderno		
date		
trackingno		
custid		

customers		
custid		
fname		
lname		
address		
city		
state		
zip		

```
{
      "orders": [
        {
              "orderno": "748745375",
              "date": "June 30, 2088 1:54:23 AM",
              "trackingno": "TN0039291"
              "custid": "11045",
              "customers": [
                   {
                        "custid": "11045",
                        "fname": "Sue",
                        "lname": "Hatfield",
                        "address": "1409 Silver Street",
                        "city": "Ashland",
                        "state": "NE",
                        "zip": "68003"
                   }
              ]
        }
      ]
}
```

Data Gathering Methods

3 Query an API (usually web-based)

This involves writing queries with the aid of a programming language to access data via an API.

Such queries are often issued via HTTP Request. That is, The ***vast majority of automated data queries*** will run via HTTP requests.

Data Gathering Methods

3 Query an API (usually web-based)

HTTP Request Basics:

HTTP GET is the most common method, but there are also **PUT**, **POST**, **DELETE** methods change some state on the server.

Data Gathering Methods

4 Scrap data from a webpage

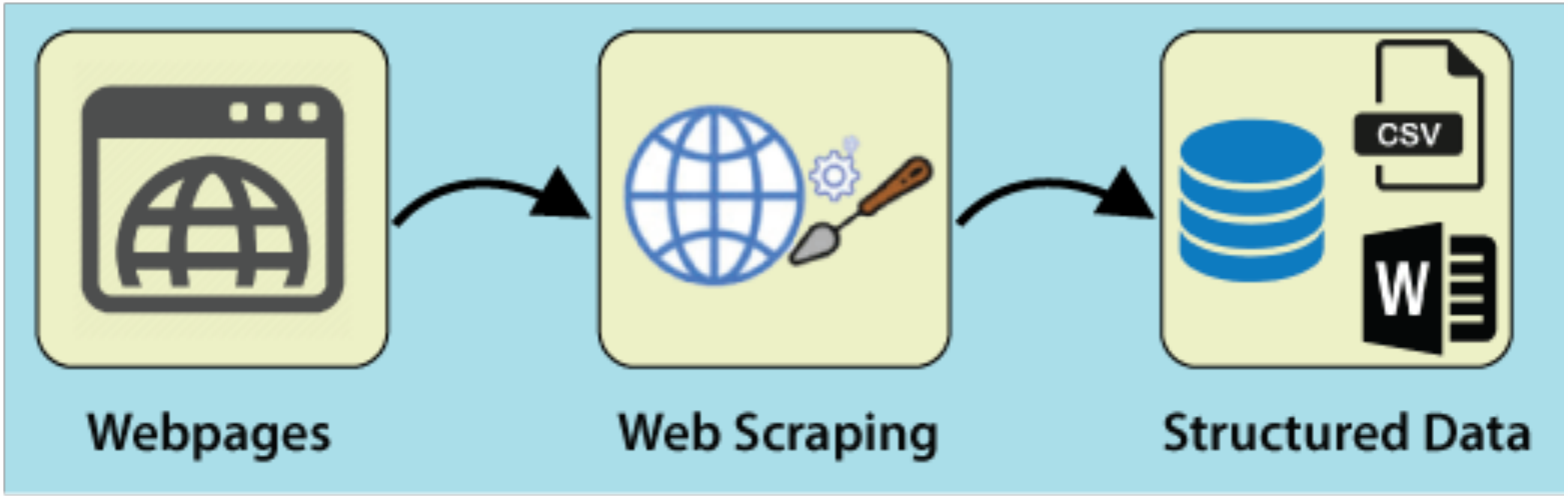
Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.

Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser.

Data Gathering Methods

4 Scrap data from a webpage

While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.



❑ Forms of Gathering Dataset for DS Project

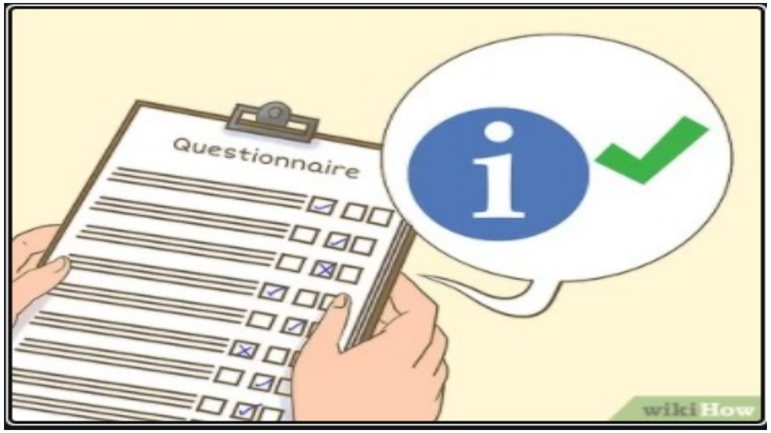
5 Acquisition of data by oneself

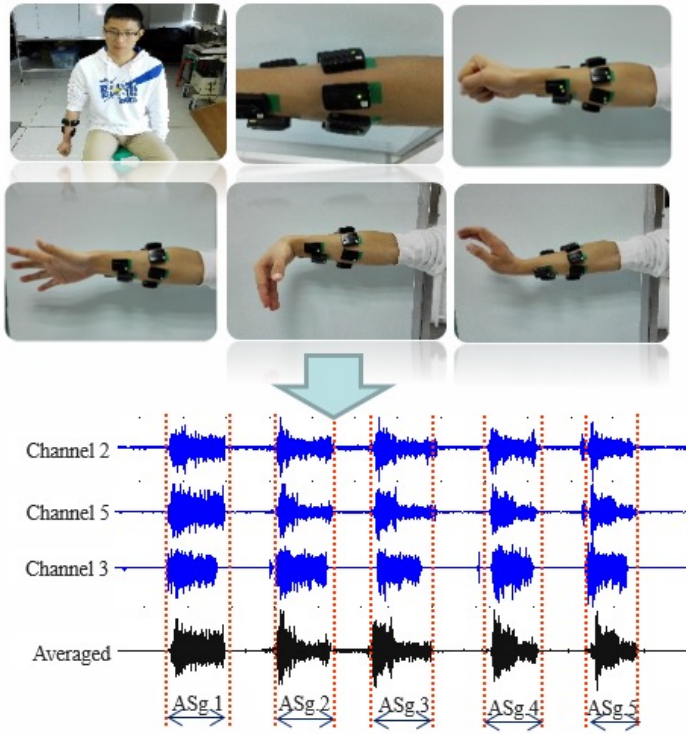
Another means of obtaining data for Data Science solution would be to collected the data based on designed experimental protocol.

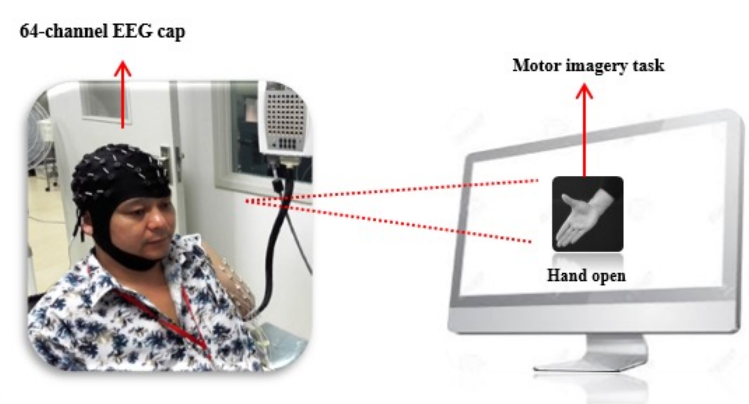
Though this approach is mostly not considered in the field of Data Science because the data gotten from such means are usually not large in volume.

❑ Forms of Gathering Dataset for DS Project

5 Acquisition of data by oneself







❑ Common Data Formats and Handling

✓ The three most common formats (judging by my completely subjective experience):

- CSV (comma separate value) files
- JSON (Javascript object notation) files and strings
- HTML/XML (hypertext markup language / extensible markup language) files
- and strings

❑ Common Data Formats and Handling

✓ CSV Files:

✓ Refers to any delimited text file (not always separated by commas)

```
"Semester","Course","Section","Lecture","Mini","Last Name","Preferred/First
Name","MI","Andrew ID","Email","College","Department","Class","Units","Grade
Option","QPA Scale","Mid-Semester Grade","Final Grade","Default Grade","Added
By","Added On","Confirmed","Waitlist Position","Waitlist Rank","Waitlisted
By","Waitlisted On","Dropped By","Dropped On","Roster As Of Date”
"F16","15688","B","Y","N",”Kolter","Zico","","zkolter","zkolter@andrew.cmu.edu","S
CS","CS","50","12.0","L","4+"," "," ","","reg","1 Jun
2016","Y","","","","","","","30 Aug 2016 4:34"
```

❑ Common Data Formats and Handling

✓ CSV Files:

If values themselves contain commas, you can enclose them in quotes

```
import pandas as pd
dataframe = pd.read_csv("CourseRoster_F16_15688_B_08.30.2016.csv", delimiter=',',
quotechar='"')
```

We’ll talk about the pandas library more in later lectures

❑ Common Data Formats and Handling

✓ JSON files / string:

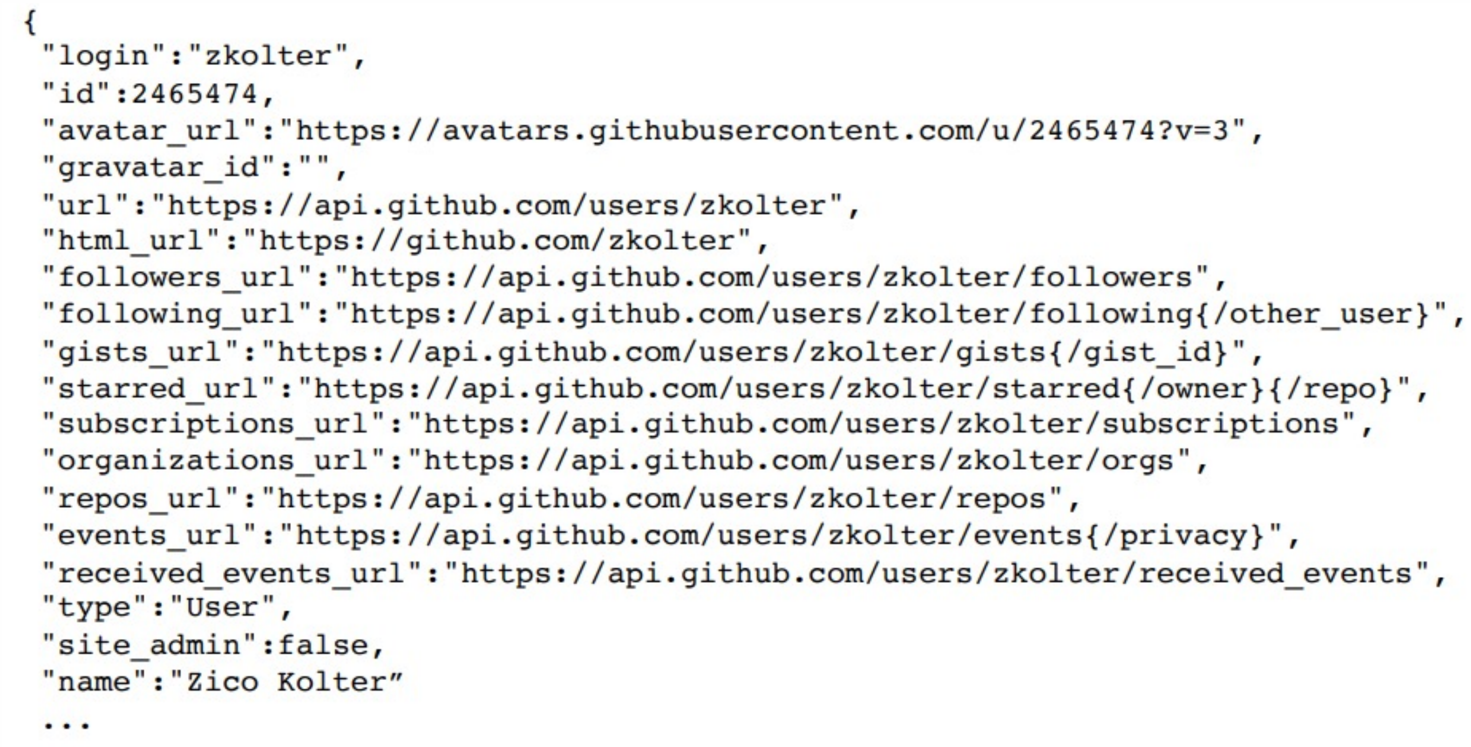
JSON originated as a way of encapsulating Javascript objects. A number of different data types can be represented

- Number: 1.0 (always assumed to be floating point)
- String: "string"
- Boolean: true or false
- List (Array): [item1, item2, item3,…]
- Dictionary (Object in Javascript): {"key":value}

❑ Common Data Formats and Handling

✓ Example JSON data:

JSON from Github API



```
{
 "login":"zkolter",
 "id":2465474,
 "avatar_url":"https://avatars.githubusercontent.com/u/2465474?v=3",
 "gravatar_id":"",
 "url":"https://api.github.com/users/zkolter",
 "html_url":"https://github.com/zkolter",
 "followers_url":"https://api.github.com/users/zkolter/followers",
 "following_url":"https://api.github.com/users/zkolter/following{/other_user}",
 "gists_url":"https://api.github.com/users/zkolter/gists{/gist_id}",
 "starred_url":"https://api.github.com/users/zkolter/starred{/owner}{/repo}",
 "subscriptions_url":"https://api.github.com/users/zkolter/subscriptions",
 "organizations_url":"https://api.github.com/users/zkolter/orgs",
 "repos_url":"https://api.github.com/users/zkolter/repos",
 "events_url":"https://api.github.com/users/zkolter/events{/privacy}",
 "received_events_url":"https://api.github.com/users/zkolter/received_events",
 "type":"User",
 "site_admin":false,
 "name":"Zico Kolter”
 ...
```

❑ Common Data Formats and Handling

✓ Regular expressions and Parsing:

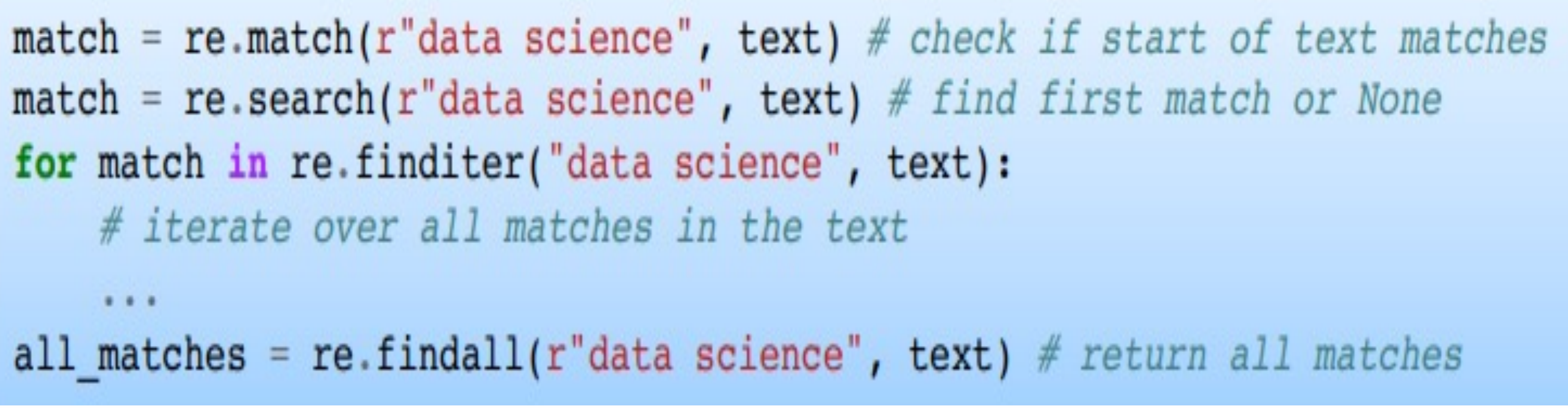
After loading the data, you will often need to search for specific elements within it.

E.g., find the first occurrence of the string “data science”

❑ Common Data Formats and Handling

✓ Regular expressions and Parsing:

Regular expressions in Python: Below are a few methods used to call Regular Expressions in Python…

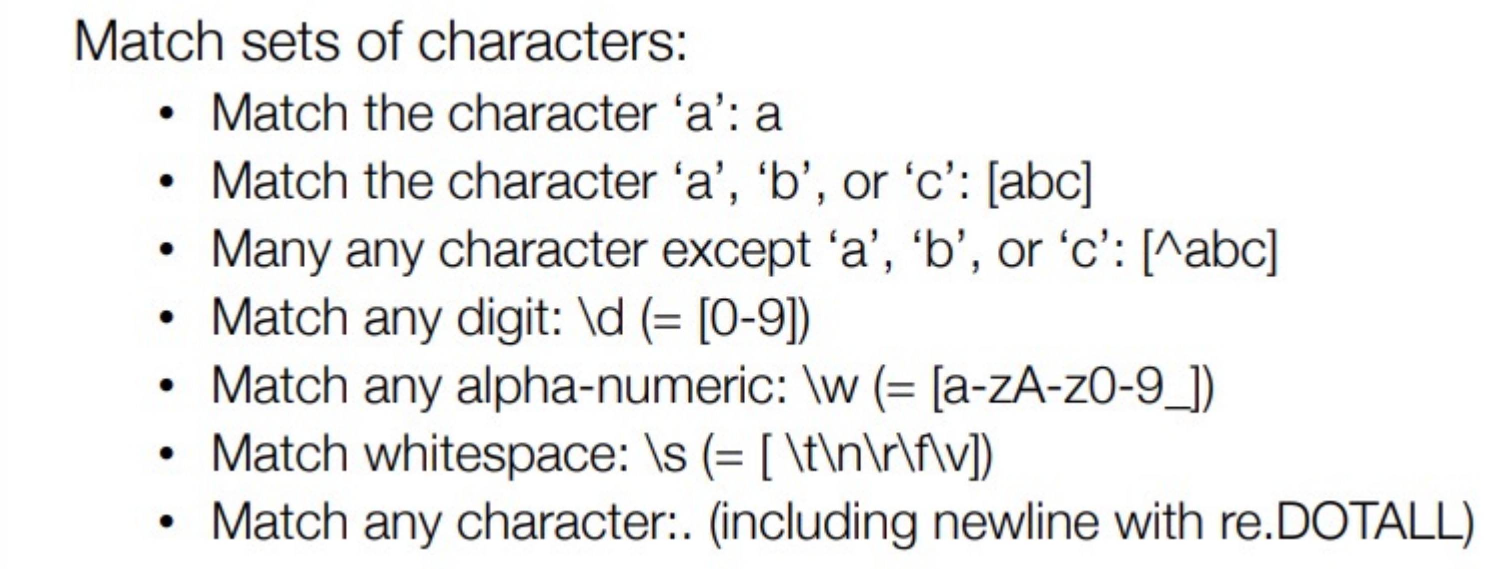


```
match = re.match(r"data science", text) # check if start of text matches
match = re.search(r"data science", text) # find first match or None
for match in re.finditer("data science", text):
    # iterate over all matches in the text
    ...
all_matches = re.findall(r"data science", text) # return all matches
```

❑ Common Data Formats and Handling

✓ Regular expressions and Parsing:

Matching multiple potential characters: The real power of REs comes in their ability to match multiple possible sequence of characters.



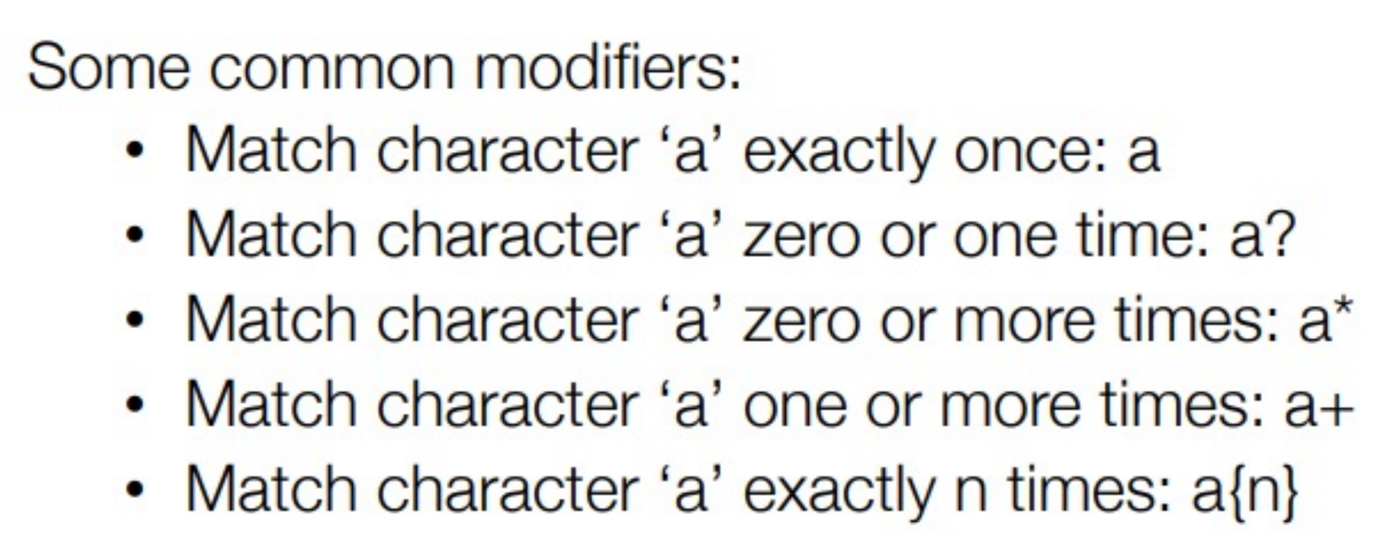
Match sets of characters:

- Match the character ‘a’: a
- Match the character ‘a’, ‘b’, or ‘c’: [abc]
- Many any character except ‘a’, ‘b’, or ‘c’: [^abc]
- Match any digit: \d (= [0-9])
- Match any alpha-numeric: \w (= [a-zA-z0-9_])
- Match whitespace: \s (= [\t\n\r\f\v])
- Match any character:. (including newline with re.DOTALL)

❑ Common Data Formats and Handling

✓ Regular expressions and Parsing:

Matching repeated characters: Can match one or more instances of a character (or set of characters)



Some common modifiers:

- Match character ‘a’ exactly once: a
- Match character ‘a’ zero or one time: a?
- Match character ‘a’ zero or more times: a*
- Match character ‘a’ one or more times: a+
- Match character ‘a’ exactly n times: a{n}

❑ Common Data Formats and Handling

✓ Regular expressions and Parsing:

Substitutions: Regular expressions provide a mechanism for replacing some text with outer text

❑ Conclusion

- ✓ There are a number of ways through which Data can be obtained for Data Science project execution.

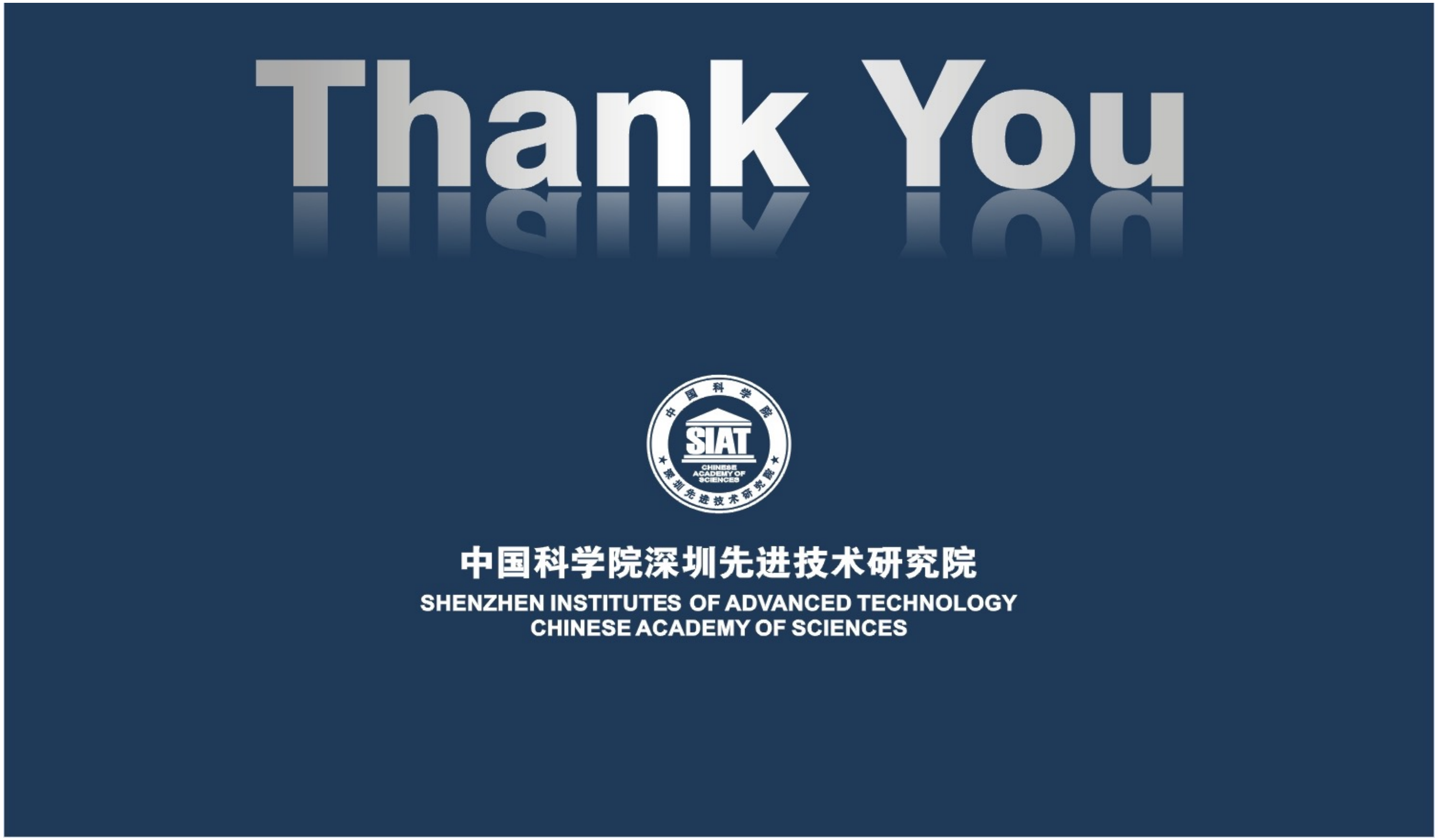
- ✓ Data Scientists should consider utilizing the best approach for collecting data

- ✓ The should also apply some simple data formatting techniques to clean up such data.

❖ Summary for today's lecture

✓ We have learned about different data collection modalities and some fundamental data formatting techniques, as required in a Data Science Project

Questions and Comments!


Thank You
中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES