THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

PolyU 理大商學院
Business School
Innovation-driven Education and Scholarship

School of
ACCOUNTING
& FINANCE
會計及金融學院

Week 6: Introduction to Scientific Computing, Web Protocols, REST API, Accessing and Processing Data

AF3214 Python Programming for Accounting and Finance

Vincent Y. Zhuang, Ph.D.
vincent.zhuang@polyu.edu.hk

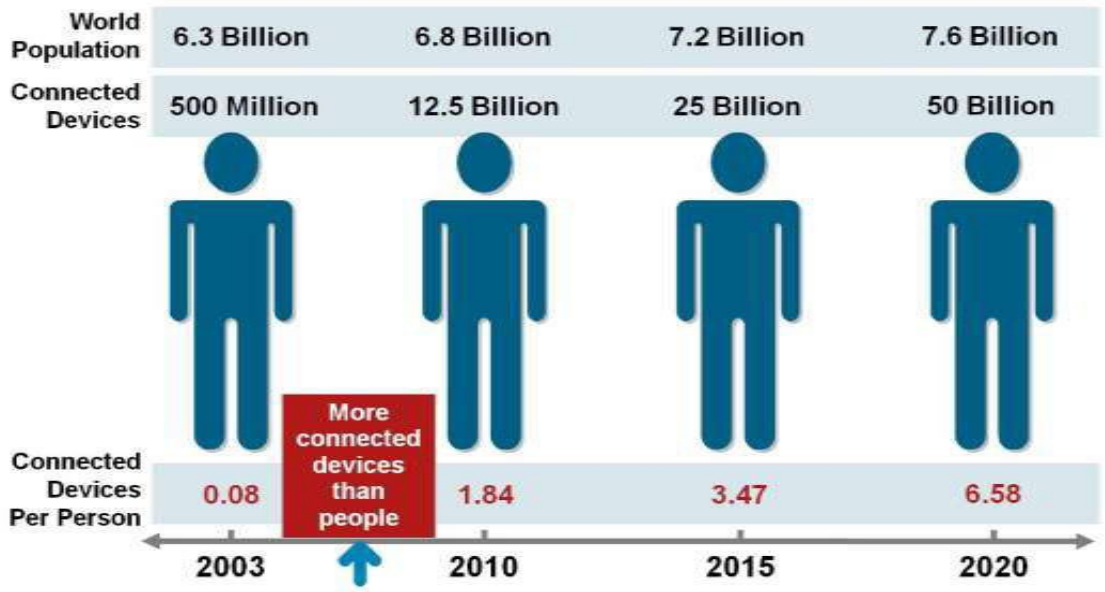
School of Accounting and Finance
The Hong Kong Polytechnic University

R508, 8:30 am – 11:20 am, Wednesdays, Semester 2, AY 2024-25

Why We need Big Data

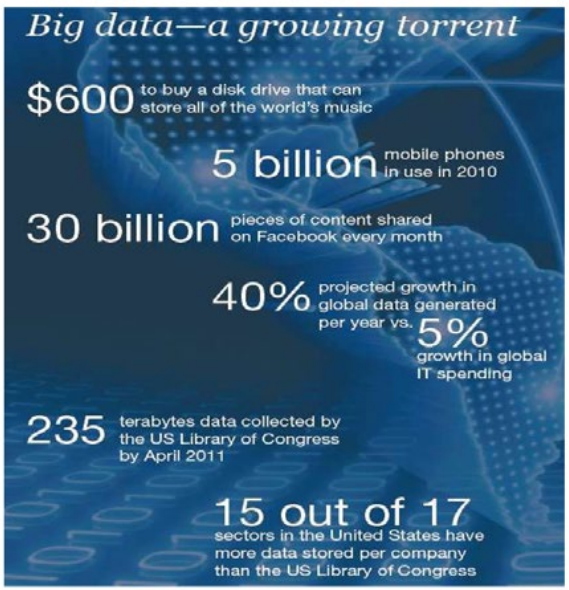
- Facebook generates 10TB of data daily
- Twitter generates 7TB of data daily
- IBM claims 90% of today's stored data was generated in just the last two years……

Figure 1. The Internet of Things Was "Born" Between 2008 and 2009



Source: Cisco IBSG, April 2011 https://www.worldometers.info/world-population/

- Mobile Devices
- Microphones
- Readers/Scanners
- Science facilities
- Programs/ Software
- Social Media
- Cameras

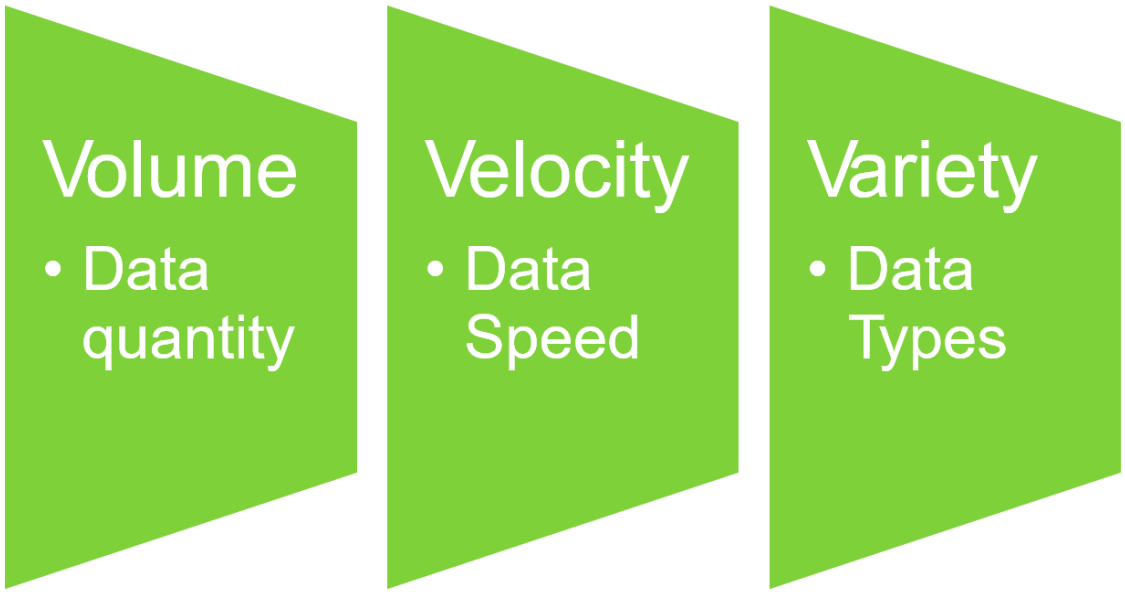


← Examples of data generation points

Big Data Characteristics

- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion posts/photos/videos from its user base.
- Decoding the human genome originally took 10 years to process; now it can be achieved in one week.

Three Characteristics Of Big Data – V3s



V3s

Volume – Data quantity

- A typical PC might have had 10 GB of storage in 2000.
- Today, Facebook has 4 petabytes (4,096 TB) of new data every day.
- Boeing 737 generates 240 TB of flight data during a single flight across The US.(Automatic Dependent Surveillance–Broadcast (ADS-B): https://www.flightradar24.com/)

Velocity – Data speed

- Clickstreams and ad impressions capture user behavior at millions of events per second.
- High-frequency trading algorithms reflect market changes in microseconds.
- Machine to machine processes exchange data between billions of devices.
- On-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

V3s

Variety – Data Types

- Big Data isn't just numbers, dates, and strings. Big Data also include location data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data.

Why We need Big Data

- Growth of Big Data is needed
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data (different data types)

Big Data Analytics

- Examining large amount of data
- Appropriate information (about data)
- Identification of hidden patterns, unknown correlations
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

Applications of Big Data Analytics

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

Search Quality

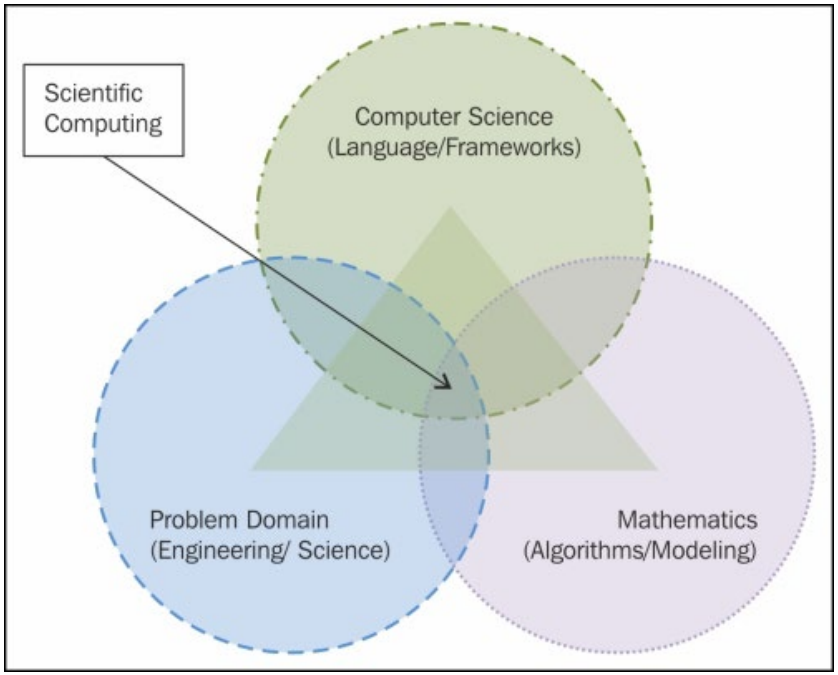
Benefits of Big Data

- Real-time big data isn't just a process for storing data in a data warehouse. It's about the ability to make better decisions and take meaningful actions at the right time.

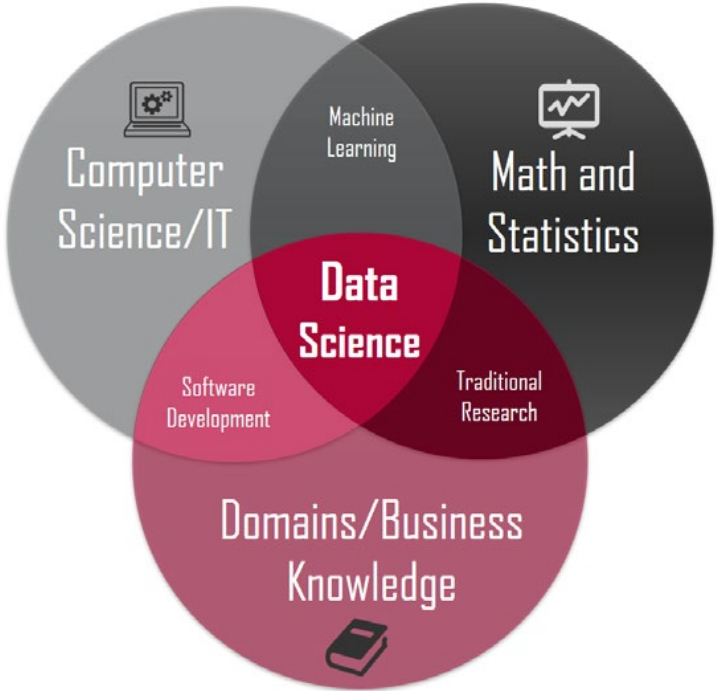
- Technologies give you the scale and flexibility to store data before you know how you are going to process it.

What are Scientific Computing and Data Science?

- "Scientific computing is the collection of tools, techniques and theories required to solve on a computer the mathematical models of problems in science and engineering."
- --Gene H. Golub and James M. Ortega



Scientific computing can be described as an interdisciplinary field, as presented in the above diagram.



Data science is the field of study that uses mathematics, programming, and domain/business knowledge to extract meaningful insights from data.

Cornerstones of The Scientific Method

Replication – Numerical calculations should be able to rerun the simulations and replicate the results upon request. Others should also be able to perform the same calculations and obtain the same results, given the information about their methods.

Reproducibility – Results obtained from an experiment, or an observational study or in a statistical analysis should be achieved again with a high degree of reliability when the study is replicated.

In other words, "Reproducibility" refers to instances in which the original researcher's data and computer codes are used to regenerate the results, while "Replicability" refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.

FinTech

FinTech? or TechFin?

- TechFin: which refers to technology firms entering the financial field
- It's basically trying to tell you the philosophical way what FinTech is and make it really kind of like it's all driven by machines
- Tech is great…but what is **not** tech? Is excel sheet tech?
- Must be something beyond our imagination. Think in a different way.
- Before embracing, better understand…

FinTech – Cont'd

My view of different parts of Tech that can help you think in the financial industry.

- **Tech** boosts quality service or strengthen one part of traditional financial industry
  - Peer-to-Peer lending on platforms (P2P)
    - Big Data → better risk control

    P2P may not be such a big industry to start with, but it's kinda like a natural supplement to the traditional financial industry
  - Insurance (fraud prevention):
    - Artificial Intelligence, Machine Learning
- **Tech** creates a new sector but not revolutionary
  - Say, robotic investment advisory based on AI and ML
- **Blockchain** is significantly different
  - Revolutionary ideas confronts fundamental (and well-studied!) finance/economics principles
- (Almost) no **tech** in ICO……

Examples of Applications in Accounting and Finance

Risk Analytics

Real-Time Analytics

Fraud Detection

Algorithmic Trading

Customer Data Management

Consumer Analytics

Providing Personalized Services

Where do we Start?

Three Common Data Types in Data Analysis

Cross-Sectional Data:

Consists of a sample data taken in a single time period with many subjects (such as individuals, firms, countries). Ordering of the data does not matter (↗, ↘, randomized order)

Time Series Data:

Consists of observations on a variable or several variables over time, or a sequence of data points indexed in time order. Ordering matters, typically presented in chronological order.

Panel (or Longitudinal) Data:

Consists of a time series for each cross-sectional member in the data. Essentially combining above two and observe the changes over a time series.

Types of Data – Cont'd

What kind of data is there in Accounting and Finance?

Quantitative Data

numeric

-Security Prices

-Income Statements

-Balance Sheets

-Forecasts

-Many Others

Qualitative Data

non-numeric

-News

-Manage Discussion and Analysis in Financial Statements (MD&A)

-Many Others

Types of Data – Cont'd

Quantitative and qualitative data can be gathered from the *same* data unit depending on whether the variable of interest is numerical or categorical

Data unit	Numeric variable	= Quantitative data	Categorical variable	= Qualitative data
A person	"*How many* children do you have?"	**3** children	"In *which country* were your children born?"	**Australia**
	"*How much* do you earn?"	$**60,000** p.a.	"*What* is your occupation?"	**Photographer**
	"*How many* hours do you work?"	**38** hours per week	"Do you work *full-time or part-time*?"	**Full-time**
A house	"*How many* square metres is the house?"	**200** square metres	"In *which city or town* is the house located?"	**Sydney**
A business	"*How many* workers are currently employed?"	**264** employees	"*What* is the *industry* of the business?"	**Retail**
A farm	"*How many* milk cows are located on the farm?	**36** cows	"*What* is the main *activity* of the farm?"	**Dairy**

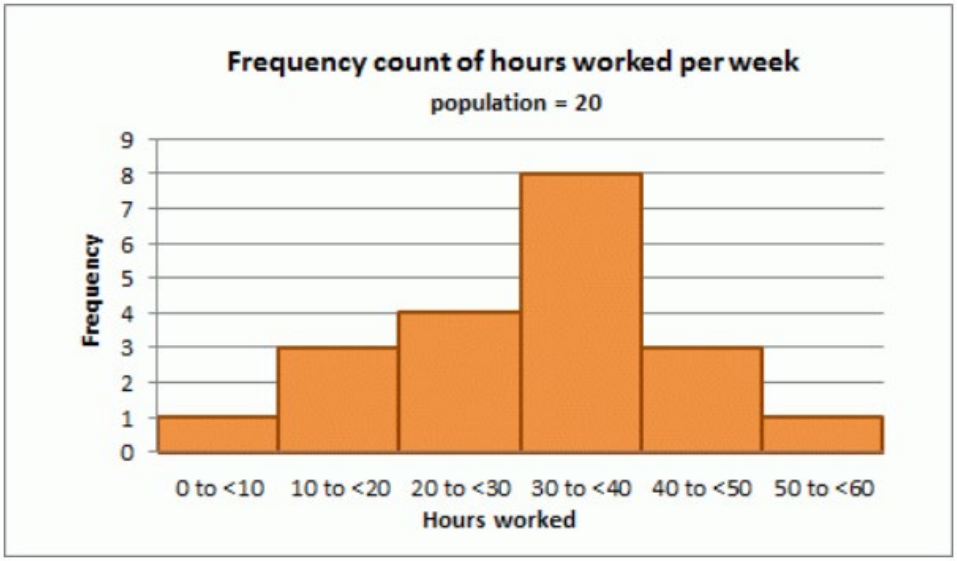
Quantitative data: Quantity, measures of values or counts and are expressed as numbers
Qualitative data: Quality, measures of 'types' and may be represented by a name, symbol, or a number code.

https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+quantitative+and+qualitative+data

Types of Data – Cont'd

The number of times an observation occurs (frequency) for a data item (variable) can be shown for both quantitative and qualitative data.

Quantitative data



Column graph showing the frequency counts of hours worked per week for 20 people.

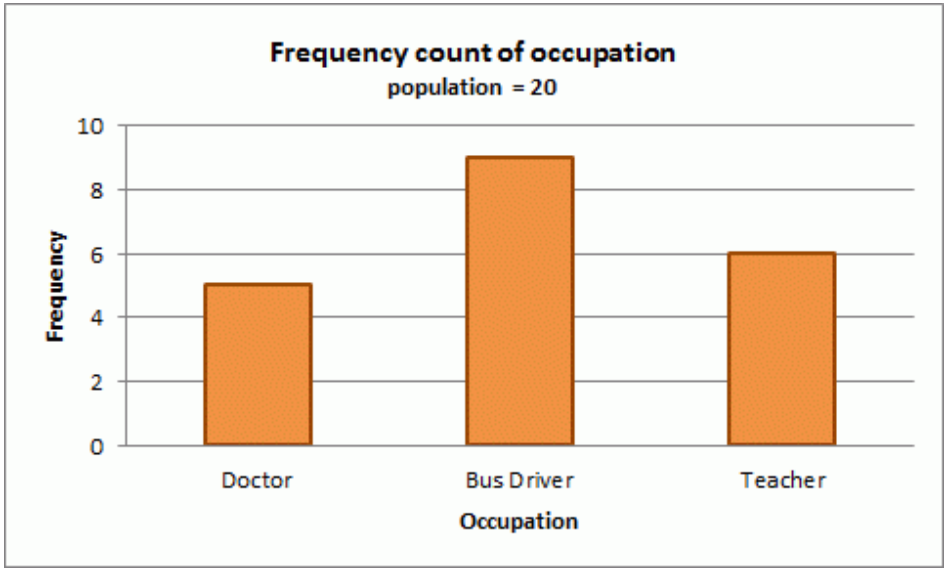
Hours worked:

0 to 10–1, 10 to 20–3,

20 to 30–4, 30 to 40–8,

40 to 50 – 3, 50 to 60 – 1.

Qualitative data



Column graph showing the frequency counts of occupation for 20 people.

Doctor - 5

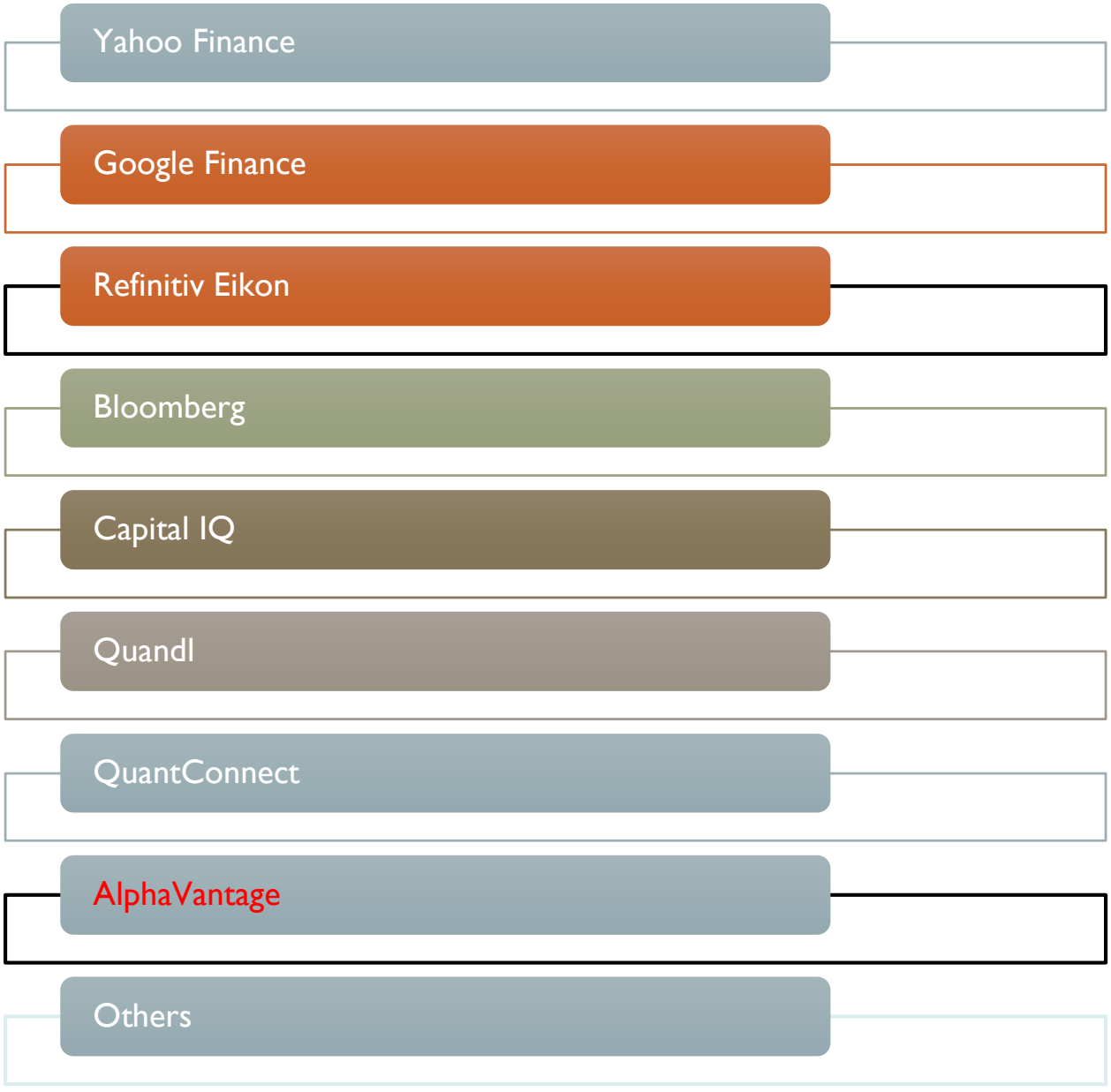
Bus driver - 9

Teacher - 6

It is important to identify whether the data are quantitative or qualitative as this affects the statistics that can be produced.

https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+quantitative+and+qualitative+data

Data Sources of Stock Price



Data Cleaning and Manipulating

You will be spending majority of your time getting data and cleaning and preparing to run through your model or program.

Understanding the structures of data is important!

Types of Structured data

HTML - Hypertext Markup Language, a standardized system for tagging text files to achieve font, colour, graphic, and hyperlink effects on World Wide Web pages.

XML - eXtensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human- and machine-readable. The World Wide Web Consortium's XML 1.0 Specification and several other related specifications—all of them free open standards—define XML. It was designed to store and transport data.

CSV - A comma-separated values file is a delimited text file that uses a comma to separate values. A CSV file stores tabular data in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas.

JSON - In computing, JavaScript Object Notation is an open-standard file format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types.

Calculating Return from Price Data

The reason for using returns versus prices is normalization. This allows us to measure all variables in a comparable metric.

It is a common practice to use log returns in finance. $P_t = P_0(1+r) = P_0 e^R \Rightarrow \ln \frac{P_t}{P_0} = \ln(1+r) = lne^R = R$

Time	Price	r	log
0	100	-	-
1	120	20.0%	18.2%
2	100	-16.7%	-18.2%
	Total Return:	3.3%?	0%

P_0 – Initial price of the stock
P_t – Price of the stock at the end of the period
R – continuously compounded rate over the period
r – simple return of the stock over time t

https://investmentcache.com/magic-of-log-returns-concept-part-1/

Descriptive/Summary Statistics

Mean/Average - The mean (or average) is the most popular and well-known measure of central tendency.

Mode - The mode is the most frequent score in our data set.

Median - The median is the middle score for a set of data that has been arranged in order of magnitude.

Descriptive/Summary Statistics - Cont'd

Variance – The averaged of the squared differences from the mean. first calculate the difference between each point and the mean; then, square and average the results.

Standard Deviation (Volatility) – Measure of how spread out numbers are. The square root of variance by figuring out the variation between each data point relative to the mean. If the points are further (closer) from the mean, there is a higher (lower) deviation.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\begin{pmatrix}\text{Standard}\\ \text{Deviation}\end{pmatrix} = \sqrt{\text{Variance}}$$

$$\sigma = \sqrt{\sigma^2}$$

S^2 = sample variance

x_i = the value of the one observation

$\bar{x}$ = the mean value of all observations

n = the number of observations

For traders and analysts, these two concepts are of paramount importance as they are used to measure security and market volatility, which in turn plays a large role in creating a profitable trading strategy.

Correlation Matrix

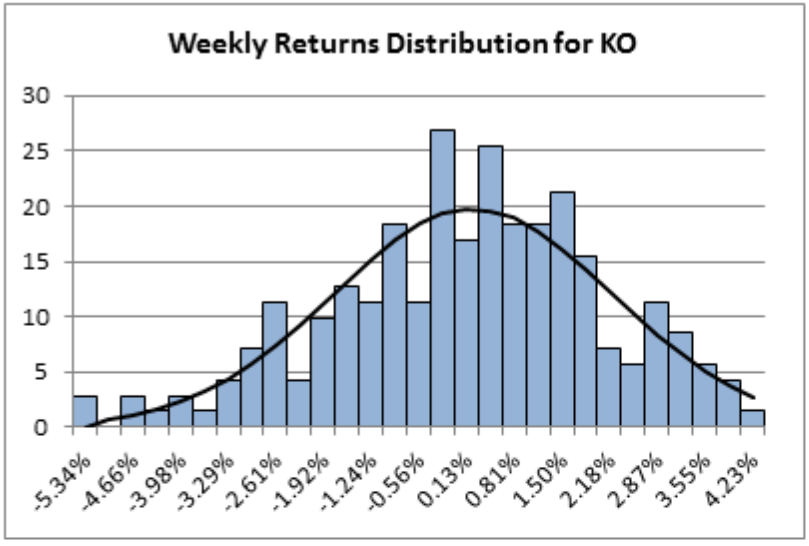
- Fundamental tool for stock market investors. It describes how closely the returns of the assets in a portfolio are correlated.
- A correlation matrix is a table showing correlation coefficients between variables.
- The line of 1.00s going from the top left to the bottom right is the main *diagonal*, which shows that each variable always perfectly correlates with itself.
- Usually, in statistics, we measure four types of correlations: *Pearson* correlation, *Kendall rank* correlation, *Spearman* correlation, and *Point-Biserial* correlation.

	iPad	iPhone	iPod	Nokia mobile phone	Other mobile phone (not Nokia and not iPhone)	Mac computer - desktop	Mac computer – laptop	PC (non-Mac)	Laptop computer (non-Mac)	None of these
iPad	1.000	.219	.143	.004	-.128	.081	.132	-.007	.009	-.027
iPhone	.219	1.000	.248	-.165	-.332	.161	.191	-.158	.152	-.053
iPod	.143	.248	1.000	-.004	-.054	.186	.238	-.012	.058	-.069
Nokia mobile phone	.004	-.165	-.004	1.000	-.423	.047	-.070	.067	-.059	-.106
Other mobile phone (not Nokia and not iPhone)	-.128	-.332	-.054	-.423	1.000	-.082	-.152	.148	.111	-.101
Mac computer - desktop	.081	.161	.186	.047	-.082	1.000	.203	-.127	.016	-.034
Mac computer – laptop	.132	.191	.238	-.070	-.152	.203	1.000	-.151	-.104	-.034
PC (non-Mac)	-.007	-.158	-.012	.067	.148	-.127	-.151	1.000	-.104	-.159
Laptop computer (non-Mac)	.009	.152	.058	-.059	.111	.016	-.104	-.104	1.000	-.148
None of these	-.027	-.053	-.069	-.106	-.101	-.034	-.034	-.159	-.148	1.000

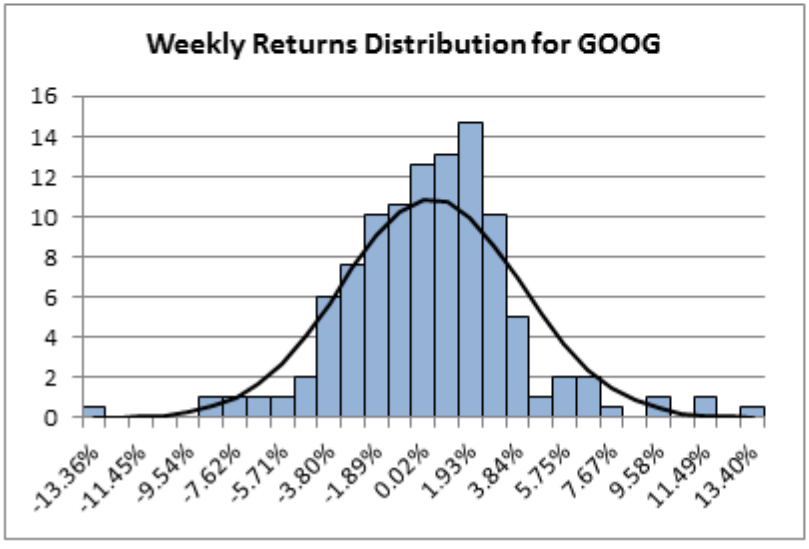
https://www.displayr.com/what-is-a-correlation-matrix/

Visualizing Data

Histograms - used to graphically summarize and display the distribution of a data set.



Mean	0.002196
Standard Deviation	0.020183
Skew	-0.21837
Kurtosis	-0.00766

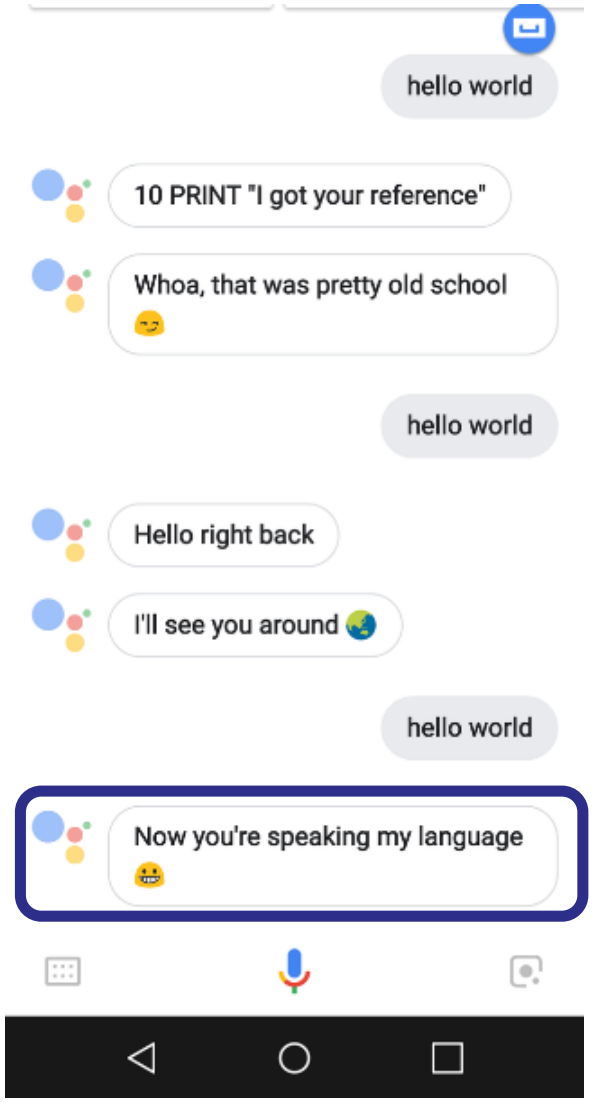


Mean	0.00363
Standard Deviation	0.036469
Skew	0.40915
Kurtosis	2.94533

Python

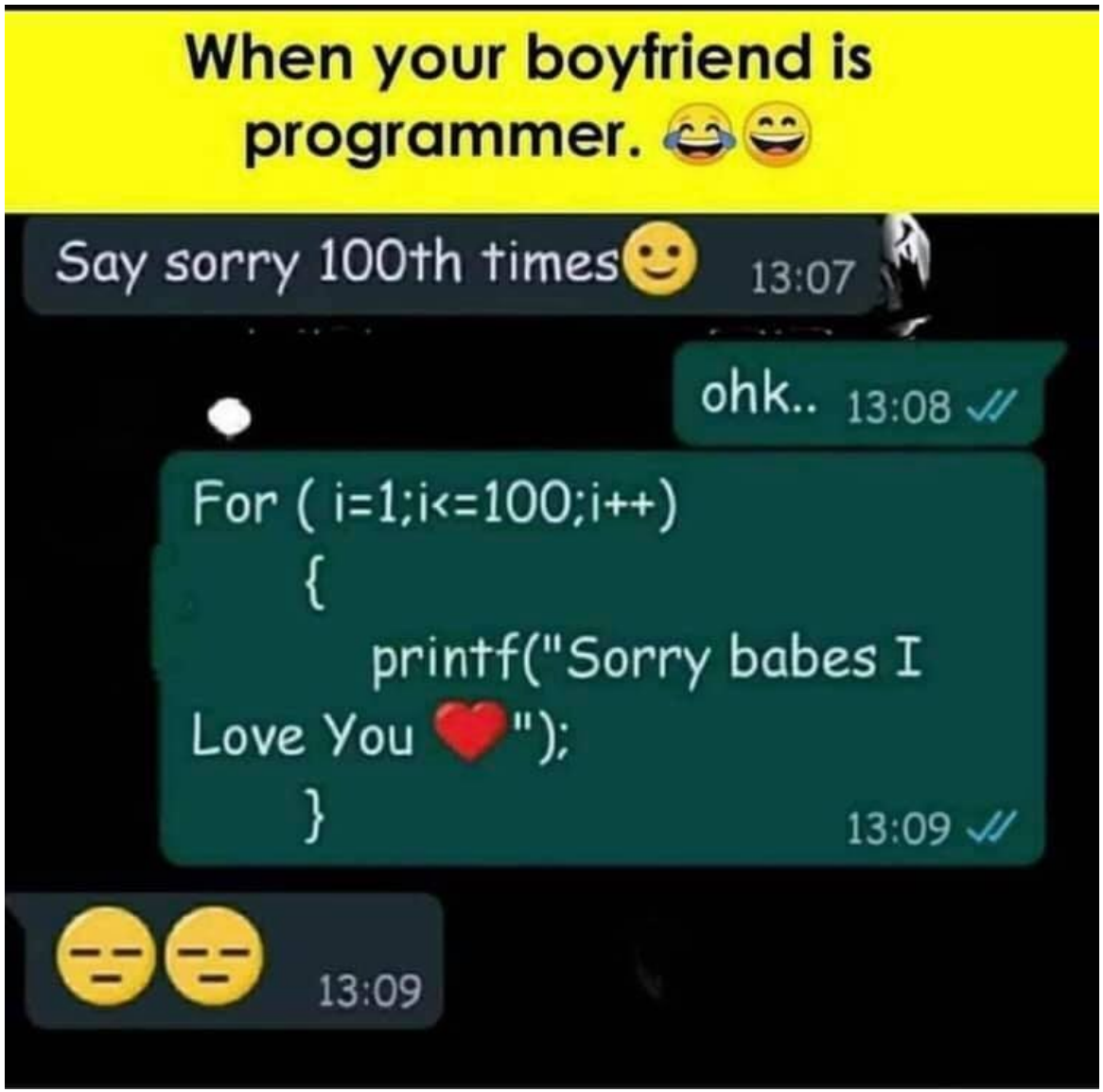
The best is yet to come…

https://www.python.org/



Python – Cont'd

Why Python?



Source: Facebook Page – I am Programmer, I have no life

Python – Cont'd

Why Python?

- It has a large community of users, easy to find help and documentation.
- Extensive ecosystem of scientific libraries and environments
- Good support for Parallel processing with processes and thread, interprocess communication, GPU computing
- Readily available and suitable for use on high-performance computing clusters

Python – Cont'd

```java
public class Main {
  public static void main(String[] args) {
    System.out.println("hello world");
  }
}
```

Why Python?

The majority of Dropbox code is written in Python, and the initial Dropbox product was almost entirely written in Python.

```
print('hello world')
```

Guido van Rossum, the creator of Python, used to work at Dropbox.

Python – Cont'd

Jupyter Notebook – An HTML-based notebook environment for Python. It is a cell-based environment with great interactivity, where calculations can be organized and documented in a structured way.

Spyder – is an IDE for scientific computing with Python. It is a powerful code editor, with syntax high-lighting, dynamic code introspection and integration with Python debugger.

IDE: An integrated development environment (IDE) is a software application that provides comprehensive facilities to computer programmers for software development.

Visual Studio Code - https://code.visualstudio.com/

An integrated development environment, supports for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

In the StackOverflow 2021 Developer Survey, Visual Studio Code was ranked the most popular developer environment tool, with 70% of 82,000 respondents reporting that they use it.^

^ https://insights.stackoverflow.com/survey/2021#section-most-popular-technologies-integrated-development-environment

Python – Cont'd

What is a Package?

A package contains all the files you need for a module.

Modules are Python code libraries you can include in your project.

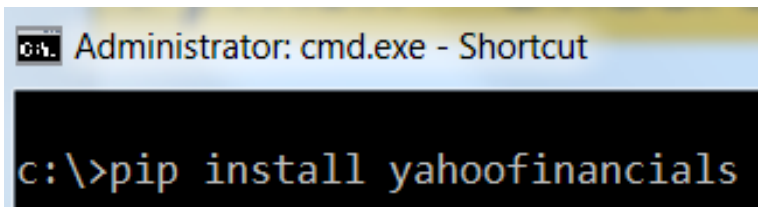
What is PIP?

PIP is a package manager for Python packages/module.

Download a Package

Open the command line interface and tell PIP to download the package you want.

Navigate your command line to the location of Python's script directory, and type the following:




```
1   import urllib
2   import urllib.request
3   from urllib.request import urlopen
4   import json
5   import time
6   from datetime import datetime
7   import requests
8   import csv
9   from urllib.parse import urlparse
10  import datetime
11  import sys
12  import io
13  import pandas as pd
```

Using a Package

https://www.w3schools.com/python/python_pip.asp, https://cs231n.github.io/python-numpy-tutorial/#numpy

Python – Understanding Data Structure

yahoofinancials

Financial data module used for pulling both fundamental and technical data from Yahoo Finance

Tesla

The module returns stock, forex, cryptocurrency, mutual fund, ETF, commodity futures, and US Treasury financial data from Yahoo Finance.

https://pypi.org/project/yahoofinancials/


```
"incomeStatementHistory": {
    "TSLA": [
        {
            "2020-12-31": {
                "researchDevelopment": 1491000000,
                "effectOfAccountingCharges": null,
                "incomeBeforeTax": 1154000000,
                "minorityInterest": 1454000000,
                "netIncome": 721000000,
                "sellingGeneralAdministrative": 3188000000,
                "grossProfit": 6630000000,
                "ebit": 1951000000,
                "operatingIncome": 1951000000,
                "otherOperatingExpenses": null,
                "interestExpense": -784000000,
                "extraordinaryItems": null,
                "nonRecurring": null,
                "otherItems": null,
                "incomeTaxExpense": 292000000,
                "totalRevenue": 31536000000,
                "totalOperatingExpenses": 29585000000,
                "costOfRevenue": 24906000000,
                "totalOtherIncomeExpenseNet": -797000000,
                "discontinuedOperations": null,
                "netIncomeFromContinuingOps": 862000000,
                "netIncomeApplicableToCommonShares": 690000000
            }
        },
```


demo.py > ...

```python
from yahoofinancials import YahooFinancials
import json
yahoo_financials = YahooFinancials('TSLA')
json_file = yahoo_financials.get_financial_stmts('annual', 'income')
with open('data.json', 'w') as f:
    json.dump(json_file, f)
```

Python – Understanding Data Structure – Cont'd

contextRef="FD2019Q2YTD" name="dei:CurrentFiscalYearEndDate" id="Fact-FD2CF7B87189BE45CF0E0F691923EF06-wk-Fact-FD2CF7B87189BE45CF0E0F691923EF06">--09-28</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:DocumentFiscalPeriodFocus" id="Fact-5150DE53B6DD5F5B812C534474C7AA63-wk-Fact-5150DE53B6DD5F5B812C534474C7AA63">Q2</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:DocumentFiscalYearFocus" id="Fact-281EFB1F9C4255C6ACE4AB4769275AF7-wk-Fact-281EFB1F9C4255C6ACE4AB4769275AF7">2019</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:EntityCentralIndexKey" id="Fact-7ED41D88286B568C85E6237D54C6C4BF-wk-Fact-7ED41D88286B568C85E6237D54C6C4BF">0000320193</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:EntityCurrentReportingStatus" id="Fact-178FCF61DF6157C7833DC5CF2DFA74F6-wk-Fact-178FCF61DF6157C7833DC5CF2DFA74F6">Yes</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:EntityEmergingGrowthCompany" id="Fact-2123243BFA345AAB83E23EE7B910B75D-wk-Fact-2123243BFA345AAB83E23EE7B910B75D" format="ixt:booleanfalse">false</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:EntityFilerCategory" id="Fact-F23C3EFA25D951DDA213AB5AF7AC2026-wk-Fact-F23C3EFA25D951DDA213AB5AF7AC2026">Large Accelerated Filer</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:EntitySmallBusiness" id="Fact-8CC7AC8D4B525DF38DF75CAEA81096EC-wk-Fact-8CC7AC8D4B525DF38DF75CAEA81096EC" format="ixt:booleanfalse">false</ix:nonNumeric><ix:nonNumeric contextRef="FD2019Q2YTD" name="dei:TradingSymbol" id="Fact-DD215FA60F77518E8675C3AD4B64BD67-wk-Fact-DD215FA60F77518E8675C3AD4B64BD67">AAPL</ix:nonNumeric><ix:nonFraction decimals="INF" unitRef="usdPerShare" contextRef="FI2018Q4" name="us-gaap:CommonStockParOrStatedValuePerShare" id="Fact-0621D55527B65679B5F11B1FEC7C547A-wk-Fact-0621D55527B65679B5F11B1FEC7C547A">0.00001</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="usdPerShare" contextRef="FI2019Q2" name="us-gaap:CommonStockParOrStatedValuePerShare" id="Fact-62A93FD278E7559C9362CB2E65A6B6ED-wk-Fact-62A93FD278E7559C9362CB2E65A6B6ED">0.00001</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="shares" contextRef="FI2018Q4" name="us-gaap:CommonStockSharesAuthorized" id="Fact-1B2BCCB06864569C9BD714368531A377-wk-Fact-1B2BCCB06864569C9BD714368531A377">12600000000</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="shares" contextRef="FI2019Q2" name="us-gaap:CommonStockSharesAuthorized" id="Fact-693FF144846155C69CA916AAFC9FB91F-wk-Fact-693FF144846155C69CA916AAFC9FB91F">12600000000</ix:nonFraction><ix:nonFraction decimals="-3" unitRef="shares" contextRef="FI2018Q4" name="us-gaap:CommonStockSharesIssued" id="Fact-C1D2AB145FBB595BB383F115673CEC27-wk-Fact-C1D2AB145FBB595BB383F115673CEC27">4754986000</ix:nonFraction><ix:nonFraction decimals="-3" unitRef="shares" contextRef="FI2019Q2" name="us-gaap:CommonStockSharesIssued" id="Fact-40073D79F5915E53AA9998404527C697-wk-Fact-40073D79F5915E53AA9998404527C697">4607284000</ix:nonFraction><ix:nonFraction decimals="-3" unitRef="shares" contextRef="FI2018Q4" name="us-gaap:CommonStockSharesOutstanding" id="Fact-34102594D18950E2A7C88B93CAA55D09-wk-Fact-34102594D18950E2A7C88B93CAA55D09">4754986000</ix:nonFraction><ix:nonFraction decimals="-3" unitRef="shares" contextRef="FI2019Q2" name="us-gaap:CommonStockSharesOutstanding" id="Fact-48F6E9AD91755E3E9BFE37F13F83C3C1-wk-Fact-48F6E9AD91755E3E9BFE37F13F83C3C1">4607284000</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2013DebtIssuanceMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-3837195B4DD5526893885AD65B368296-wk-Fact-3837195B4DD5526893885AD65B368296">17000000000</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2014DebtIssuanceMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-E1E6444466BB95924AF2E2CC40008ACBC-wk-Fact-E1E6444466BB95924AF2E2CC40008ACBC">12000000000</ix:nonFraction><ix:nonFraction decimals="-8" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2015DebtIssuancesMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-17FE7B6E132F5801AFFB02509CB232AF-wk-Fact-17FE7B6E132F5801AFFB02509CB232AF">27300000000</ix:nonFraction><ix:nonFraction decimals="-8" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2016DebtIssuancesMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-0EC516F3AB505E9488F6DE3439CBF568-wk-Fact-0EC516F3AB505E9488F6DE3439CBF568">24900000000</ix:nonFraction><ix:nonFraction decimals="-8" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2017DebtIssuancesMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-5EF009326CBD56AAAC5839C7ED9A13CD-wk-Fact-5EF009326CBD56AAAC5839C7ED9A13CD">28700000000</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="usd" contextRef="FI2019Q2_us-gaap_DebtInstrumentAxis_aapl_A2018DebtIssuanceMember" name="us-gaap:DebtInstrumentFaceAmount" id="Fact-9DD31F14219755F1841DEAFCC85128CD-wk-Fact-9DD31F14219755F1841DEAFCC85128CD">7000000000</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2013DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom2.400to3.850NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-9292BD9070165829 8D8EE6B13CA8B063-wk-Fact-9292BD9070165829 8D8EE6B13CA8B063">0.0385</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2014DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom2.100to4.450NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-616FFB2426205515B14BD104D7812EAA-wk-Fact-616FFB2426205515B14BD104D7812EAA">0.0445</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2015DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom0.350to4.375NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-FA657FE8307E5118975D1E5E592361E5-wk-Fact-FA657FE8307E5118975D1E5E592361E5">0.04375</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2016DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom1.100to4.650NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-8680796FD60456408FEF55C8BD3FC407-wk-Fact-8680796FD60456408FEF55C8BD3FC407">0.0465</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2017DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom0.875to4.300NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-3460C2C991D65FB2820DC165CA139FE0-wk-Fact-3460C2C991D65FB2820DC165CA139FE0">0.043</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MaximumMember_us-gaap_DebtInstrumentAxis_aapl_A2018DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom1.800to3.750NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-DC24D8A337D95C2AA23926AC9B55E6F2-wk-Fact-DC24D8A337D95C2AA23926AC9B55E6F2">0.0375</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2013DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom2.400to3.850NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-A521AE36827551D18907386FEE72BB64-wk-Fact-A521AE36827551D18907386FEE72BB64">0.024</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2014DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom2.100to4.450NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-A11AD4DA3AA45B0FAF426F3250595790-wk-Fact-A11AD4DA3AA45B0FAF426F3250595790">0.021</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2015DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom0.350to4.375NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-39B4C0E00AF75BC1B533D4241B31B0A6-wk-Fact-39B4C0E00AF75BC1B533D4241B31B0A6">0.0035</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2016DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom1.100to4.650NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-BFB7B4D715A350AB92516382C65F7956-wk-Fact-BFB7B4D715A350AB92516382C65F7956">0.011</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2017DebtIssuancesMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom0.875to4.300NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-115B06B6C92B5078870BC760B54F3C13-wk-Fact-115B06B6C92B5078870BC760B54F3C13">0.00875</ix:nonFraction><ix:nonFraction decimals="INF" unitRef="number" contextRef="FI2019Q2_srt_RangeAxis_srt_MinimumMember_us-gaap_DebtInstrumentAxis_aapl_A2018DebtIssuanceMember_us-gaap_LongtermDebtTypeAxis_aapl_FixedRateRangefrom1.800to3.750NotesMember" name="us-gaap:DebtInstrumentInterestRateStatedPercentage" id="Fact-9FD3C51BC8185A1D9DE5EF5353C37D1C-wk-Fact-9FD3C51BC8185A1D9DE5EF5353C37D1C">0.018</ix:nonFraction><ix:nonNumeric contextRef="FI2019Q2_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_domain_2019-03-31" name="us-gaap:RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionPeriod1" id="Fact-E2BC913B6AF65AFEB01EB2018A2D3F84-wk-Fact-E2BC913B6AF65AFEB01EB2018A2D3F84">P1Y</ix:nonNumeric><ix:nonNumeric contextRef="FI2019Q2_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_domain_2020-03-29" name="us-gaap:RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionPeriod1" id="Fact-B42F884552FA5AD383A4E0AACBD83CA0-wk-Fact-B42F884552FA5AD383A4E0AACBD83CA0">P1Y</ix:nonNumeric><ix:nonNumeric contextRef="FI2019Q2_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_us-gaap_RevenueRemainingPerformanceObligationExpectedTimingOfSatisfactionStartDateAxis_domain_2021-03-28" name="us-

Example of HTML Data from SEC EDGAR

[Apple Inc. Filing]

How to read it in Python?

Python Modules

- Pandas - *Pandas* is a library providing high-performance, easy-to-use data structures and data analysis tools for Python programming language.
- NumPy - *NumPy* is the fundamental package for scientific computing with Python. It offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

Pandas Tutorial:
https://www.kaggle.com/learn/pandas

Numpy Library Lookup:
https://cs231n.github.io/python-numpy-tutorial/#numpy

Python Modules – Cont'd

- Scikit-learn is a free machine learning library for Python programming. It features various classification, regression and clustering algorithms.
- Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

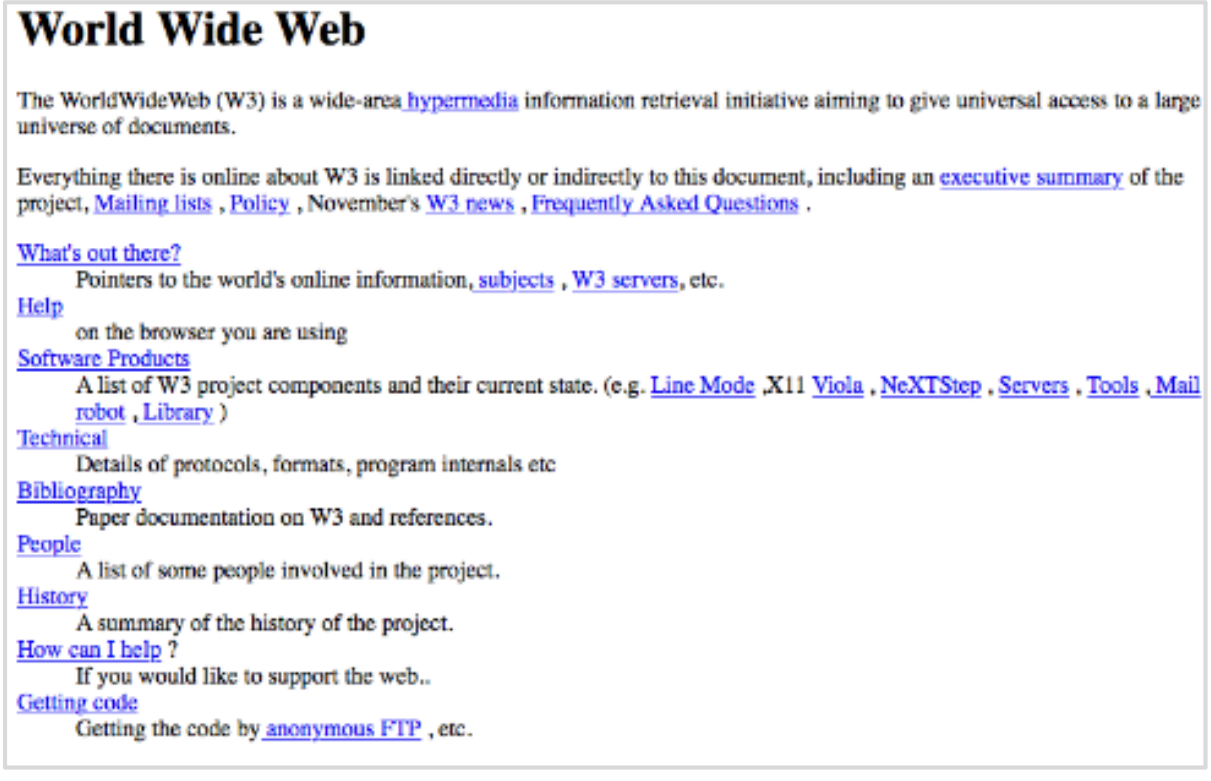
Scikit-learn Tutorial:
https://www.datacamp.com/community/tutorials/scikit-learn-python

Matplotlib Tutorial:
https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python

Let’s move to Jupyter Notebook Demo

What are Internet and WWW?

- On August 6th 1991 the very first website went online. Marked the birth of the world wide web and technological revolution.
- It was the beginning of one of the biggest achievements of the 20th century - the world wide web: http://info.cern.ch/hypertext/WWW/TheProject.html
- The man behind it is computer scientist Sir Tim Berners-Lee. In 1980, he worked as an independent contractor for CERN - The European Council for Nuclear Research.

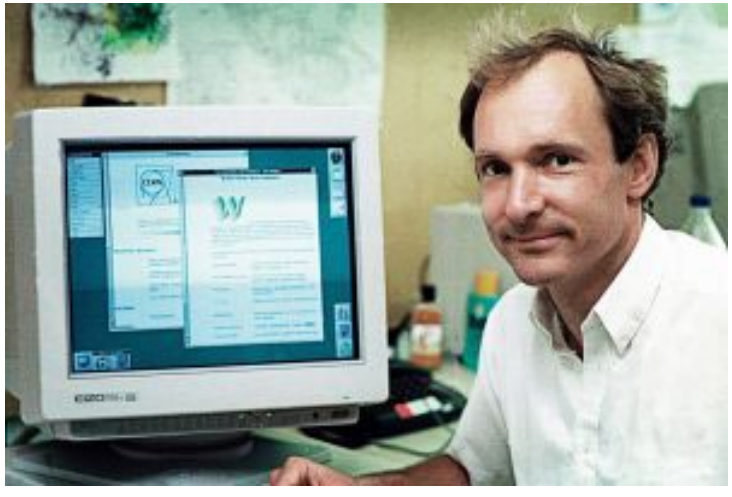


World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.



What are Internet and WWW?

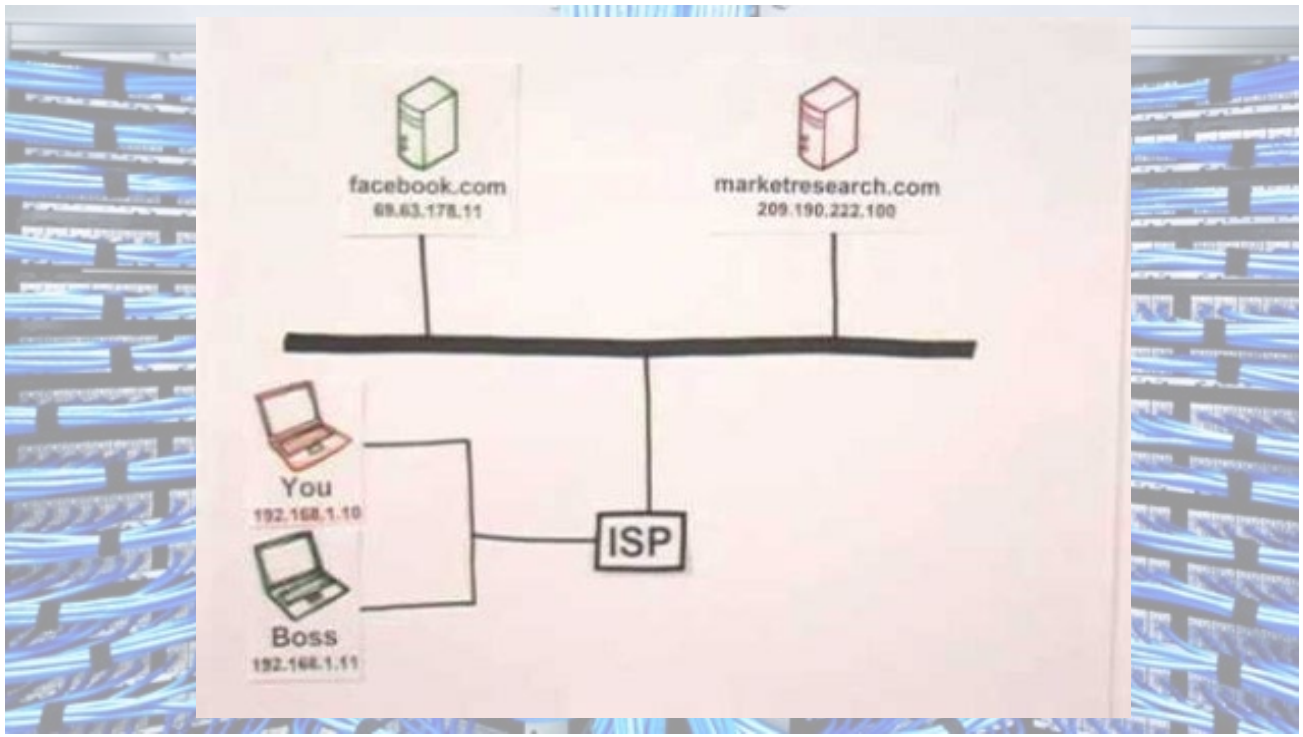
What are Internet and WWW?

What's the difference between the Web and Internet?

Web = Internet?

How does the Internet Work?

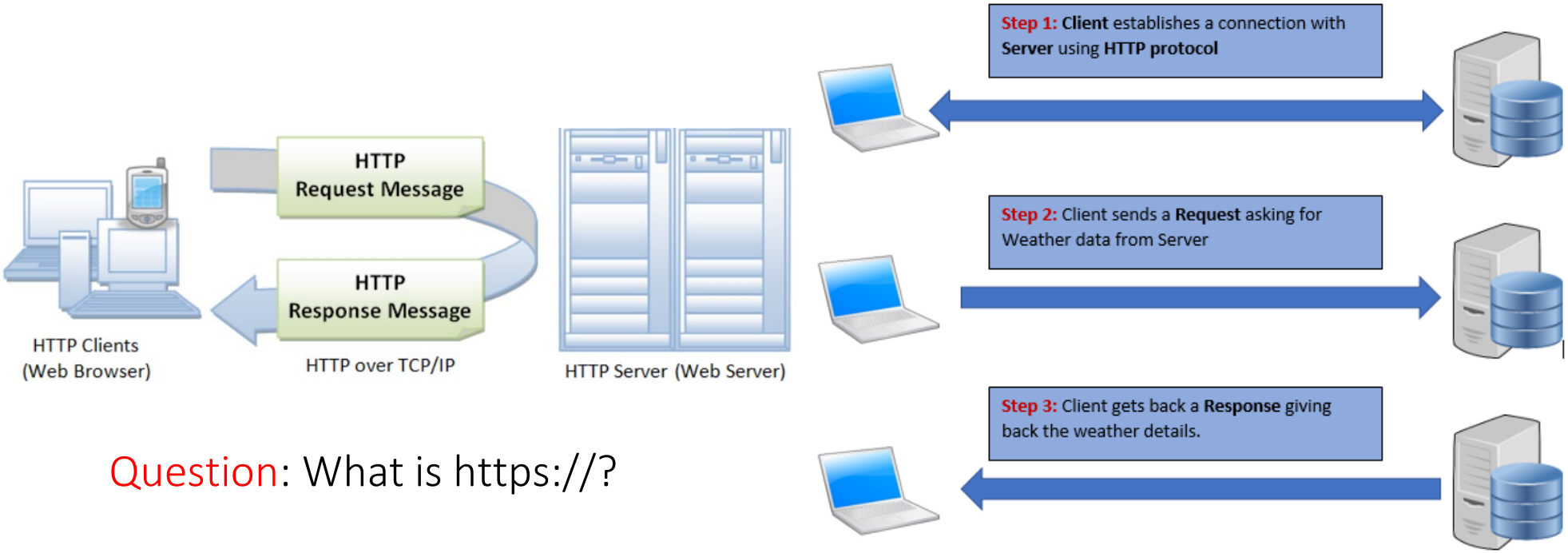
- We usually don't have to think too much about what is happening on our browser when we are surfing the internet.
- However it is important to understand the mechanisms of the Internet when we are surfing the internet for data.
- We need to understand what is happening at the browser level, as well as the network level.



How the Internet Works in 5 Minutes: https://www.youtube.com/watch?v=7_LPdttKXPc

HTTP Protocol

- HTTP (**H**yper**T**ext **T**ransfer **P**rotocol): The standard protocol for transferring web pages and content across the internet.
- When you browse a web page, the URL might be preceded by http://. This is telling the web browser to use HTTP to transfer the data.
- Web Browser is the Client and the request are sent to the servers.



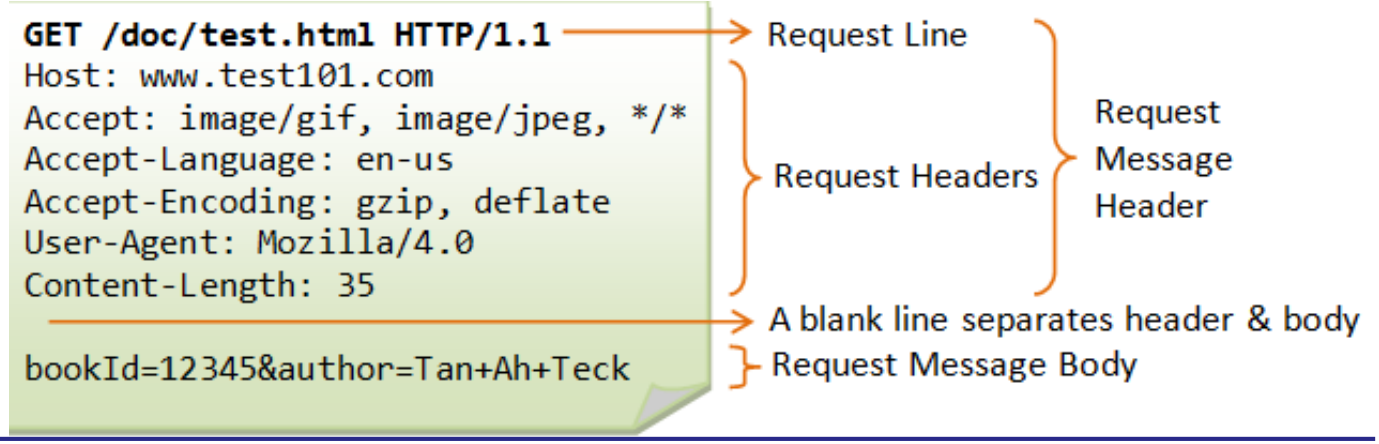
Question: What is https://?

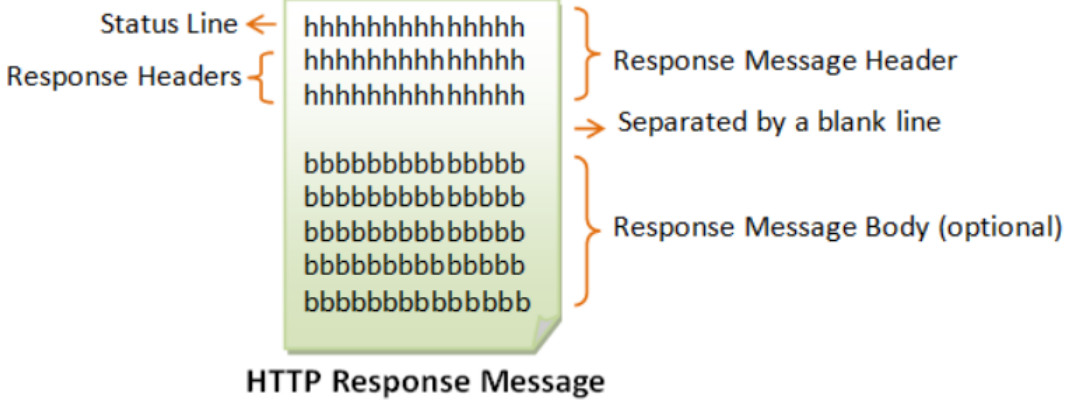
More reading: https://www.cloudflare.com/learning/ssl/what-is-https/

HTTP Request

- HTTP Request is a packet of information that one computer sends to another computer to communicate something.
- A HTTP Request contains following parts.
  - Request Line
  - Headers
  - Optional Body of the Request

HTTP request message



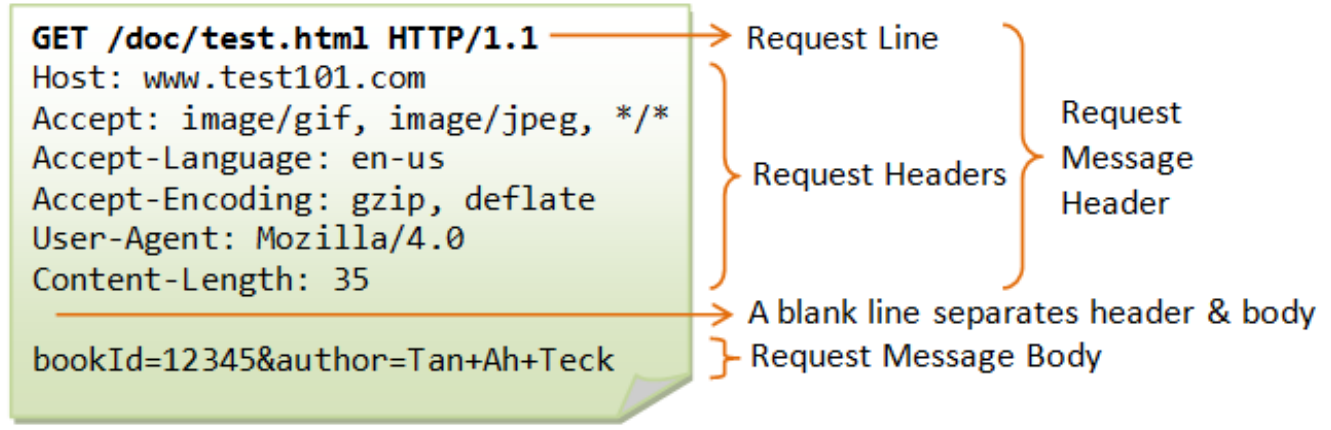


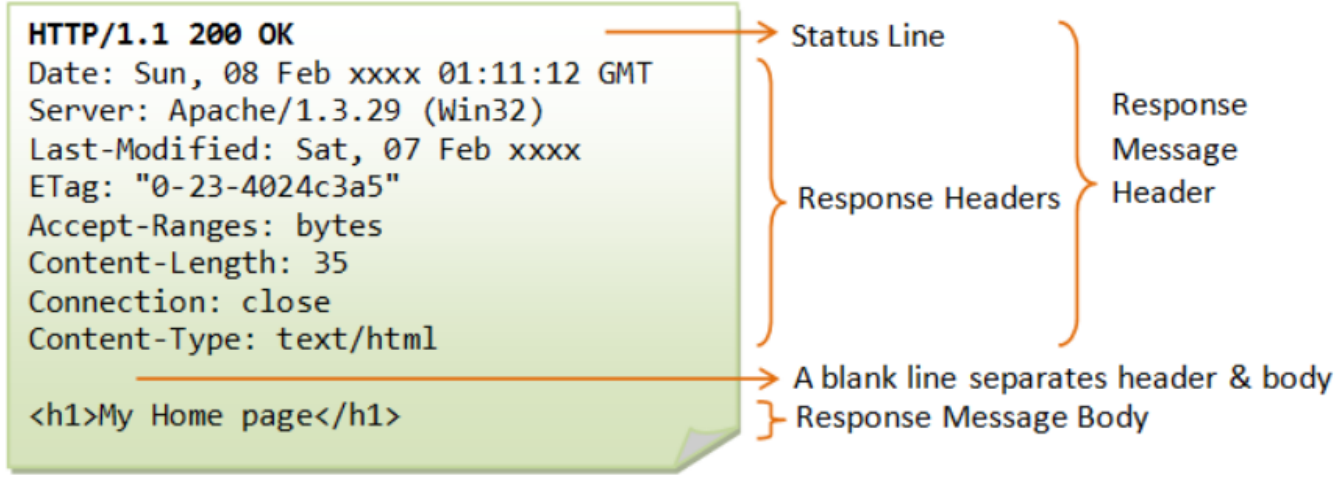
HTTP response message

The response headers are in the form: `name:value` pairs

HTTP Request – Cont'd

- It specifies the method token:
  - GET
  - POST
  - PUT
  - HEAD
  - DELETE
  - TRACE
  - OPTIONS

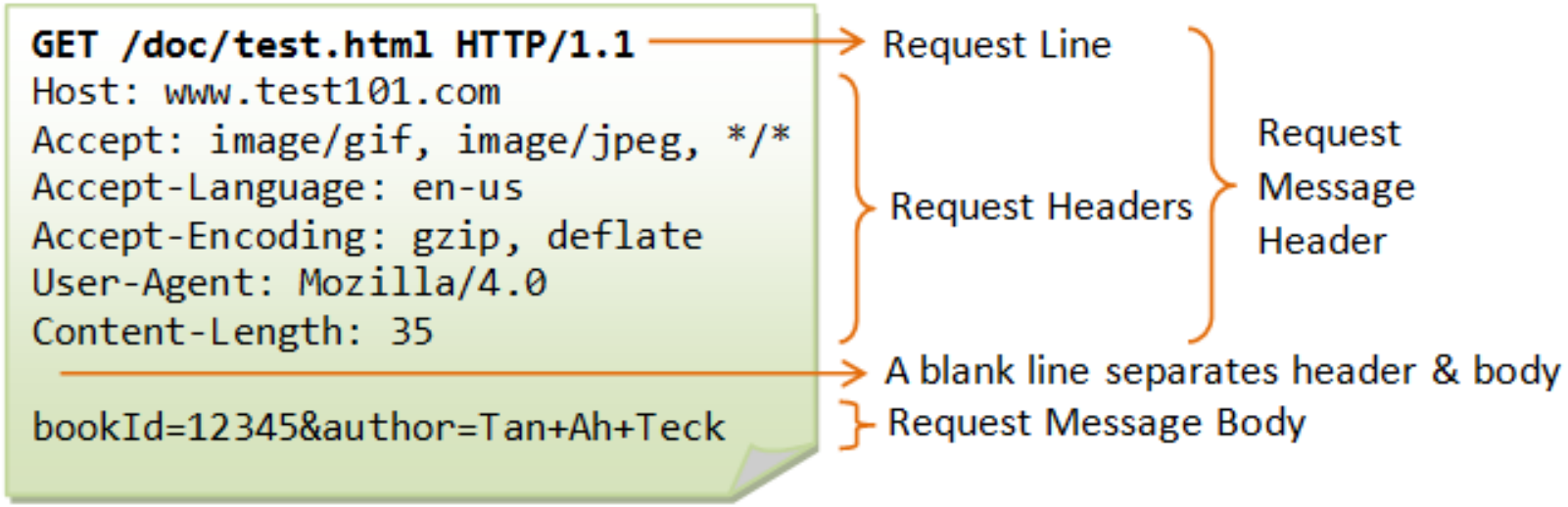




More reading: https://www.w3.org/Protocols/rfc2616/rfc2616-sec9.html

HTTP Request – Cont'd

- GET is issued to request data from a specified resource.
- It is one of the most common HTTP Methods.
- It can be cached and remains in browser history
- It should not be used when dealing with sensitive data.
- There is a length restriction – max 2,048 characters.



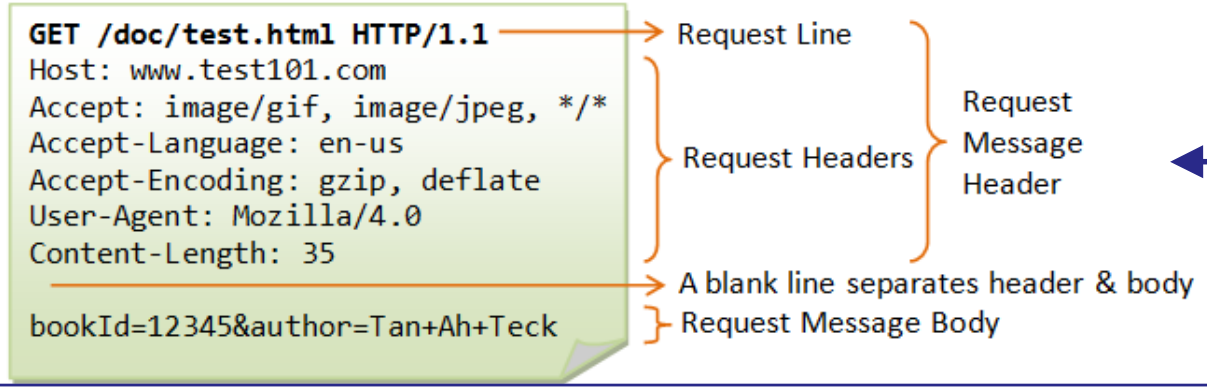
HTTP Request – Cont'd

- If you are using the GET method, you are limited to a maximum of 2,048 characters.
- However, the POST method is not limited by the size because they are transferred in the header and not in the URL.

- POST is used to send data to a server to create/update a resource.
- The data sent to the server with POST is stored in the request body.
- POST requests are never cached
- POST requests do not remain in the browser history.
- Has no restrictions on data length.

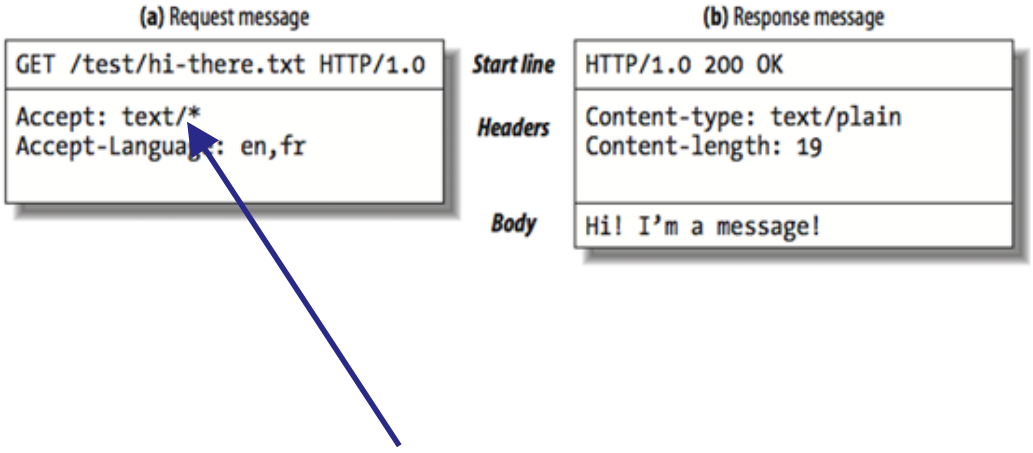
HTTP Request - Cont'd

- HTTP Request may contain zero or more Request Headers.
- In between the Request Line and the Message body is considered as Header.
- An HTTP header consists of its case-insensitive name followed by a colon (:), then by its value.
- Header is used to pass additional information about the request to the server.

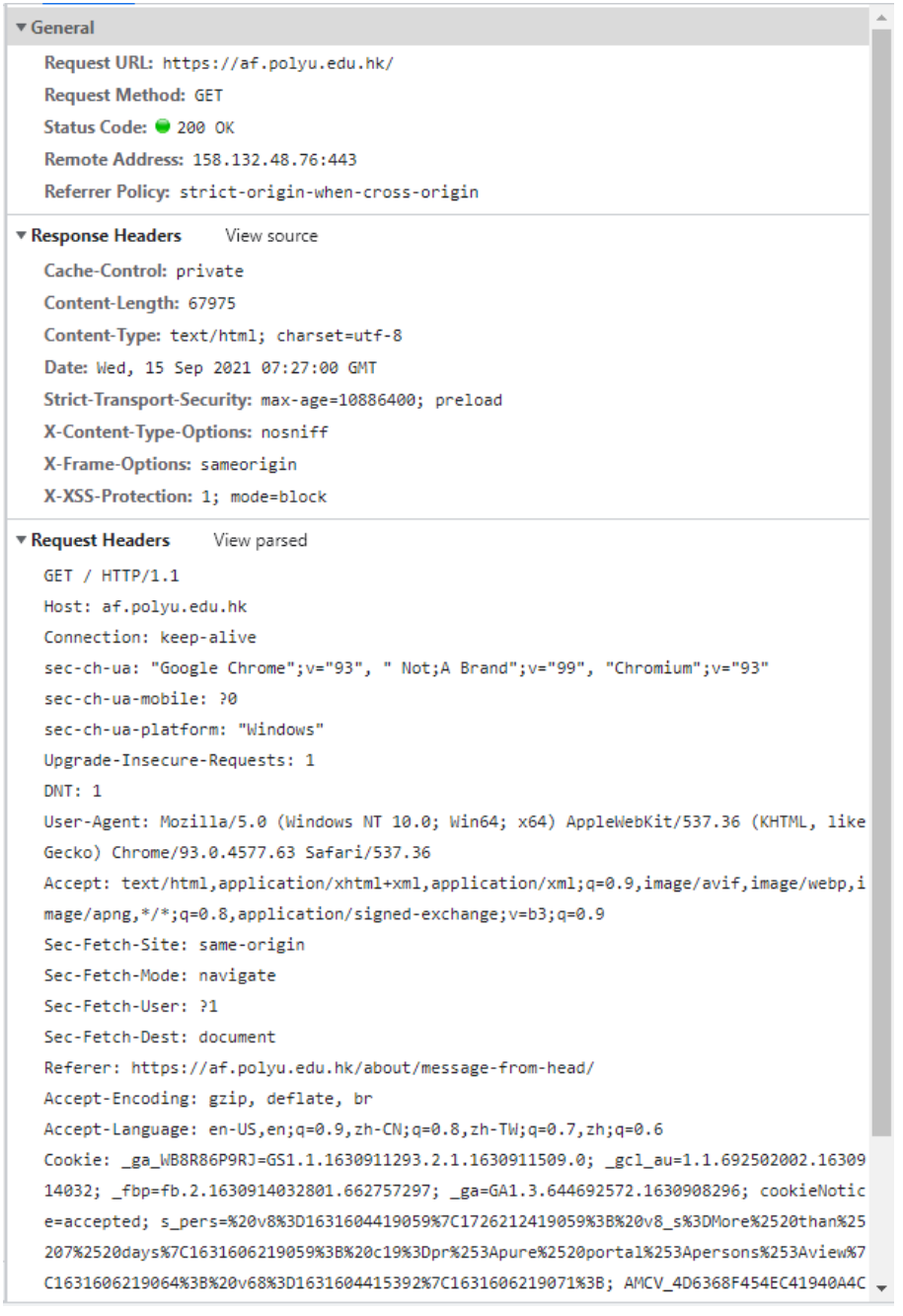


HTTP Request – Cont'd

- Request message body is the part of the HTTP request where additional content can be sent to the server.
- Request message body tries to send additional information required by the server to process current request properly.



Question: What is text/*?



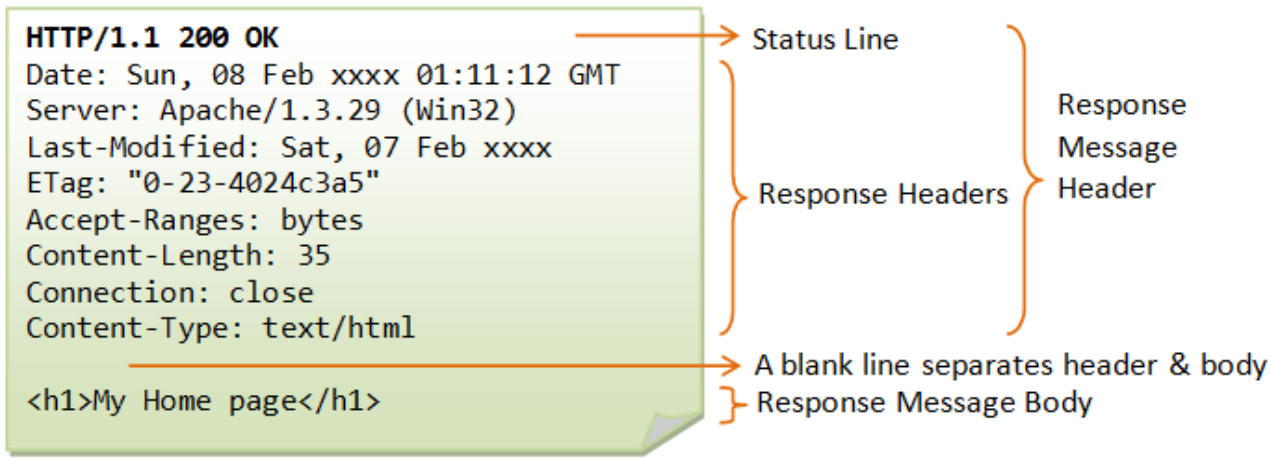
HTTP Response

After receiving and interpreting a request message, a server responds with an HTTP response message.

- The status line consists of 3 parts
  - HTTP Protocol Version
  - Status Code
  - Reason Phrase: is intended to give a short textual description of the Status Code.

Examples of status line are:

```
HTTP/1.1 200 OK
HTTP/1.0 404 Not Found
HTTP/1.1 403 Forbidden
HTTP/1.1 500 Internal
              Server Error
HTTP/1.1 502 Bad Gateway
```



More reading: https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

HTTP Response: Status Line

100	Continue
101	Switching Protocols
102	Processing

2XX Success	
200	OK
201	Created
202	Accepted
203	Non-authoritative Information
204	No Content
205	Reset Content
206	Partial Content
207	Multi-Status
208	Already Reported
226	IM Used

3XX Redirectional	
300	Multiple Choices
301	Moved Permanently
302	Found
303	See Other
304	Not Modified
305	Use Proxy
307	Temporary Redirect
308	Permanent Redirect

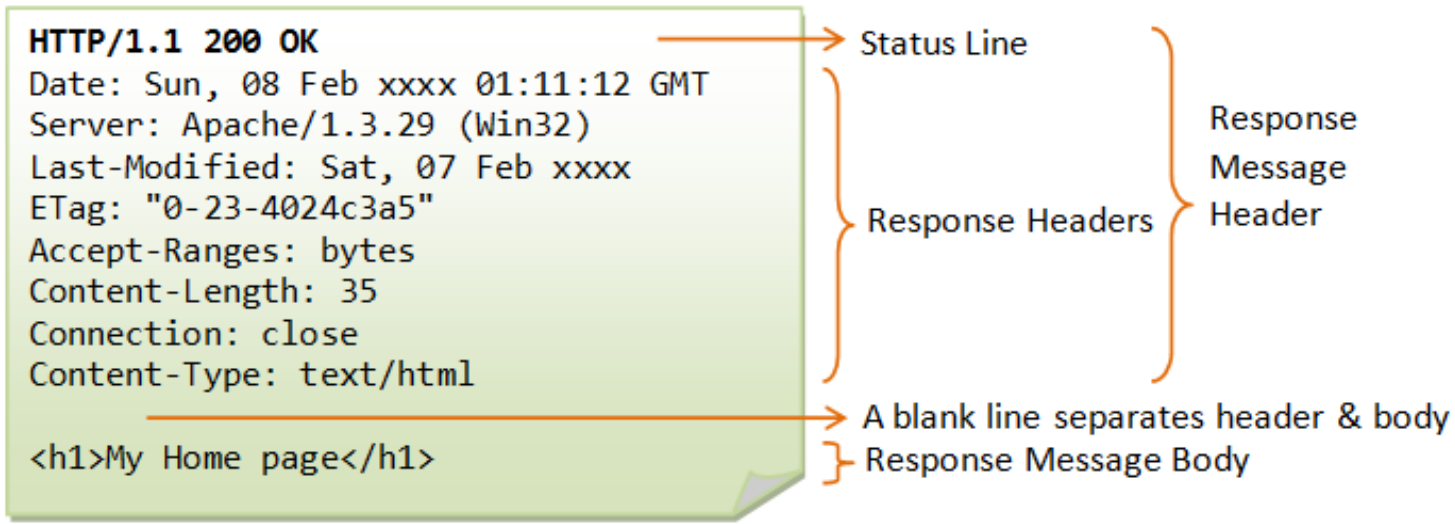
4XX Client Error	
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Payload Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
418	I'm a teapot
421	Misdirected Request
422	Unprocessable Entity
423	Locked
424	Failed Dependency
426	Upgrade Required
428	Precondition Required
429	Too Many Requests
431	Request Header Fields Too Large
444	Connection Closed Without Response
451	Unavailable For Legal Reasons
499	Client Closed Request

5XX Server Error	
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported
506	Variant Also Negotiates
507	Insufficient Storage
508	Loop Detected
510	Not Extended
511	Network Authentication Required
599	Network Connect Timeout Error

More reading: https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

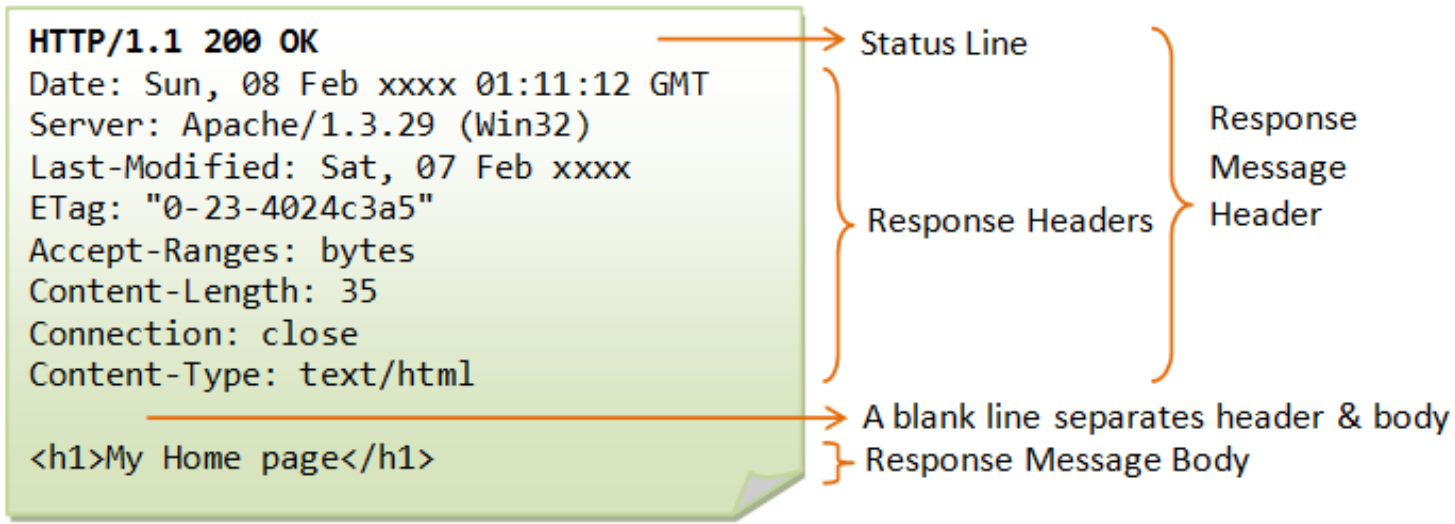
HTTP Response: Response Header

- Just like the Request header, it can contain zero or more lines.
- Very uncommon to have zero headers in the response.
- In Response header, there is a header named content-type. It is used to inform the client that body of response is a certain type.



HTTP Response: Response Body

- Contains the information requested.
- In terms of Web Services, the information requested by a client is referred to as a resource.



Request and Python

It is possible to use Python to send these HTTP Request and receive the HTTP Response.

There are many libraries that will help you achieve this goal, but we will concentrate on simplicity to start off.

Python Request library is best for making simple web requests.

https://pypi.org/project/requests/

We will have a try later on Jupyter Notebook.

Beautiful Soup

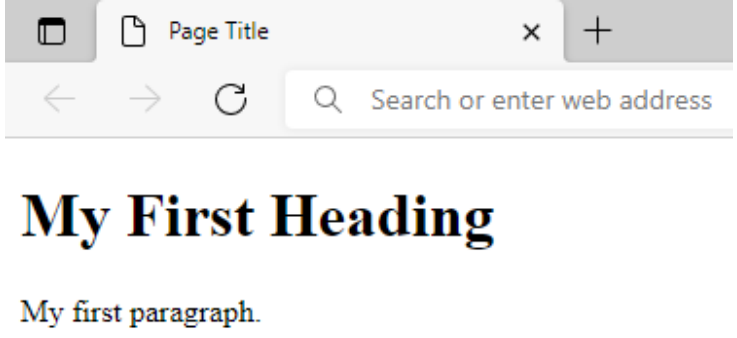
Processing HTML With Beautiful Soup 4

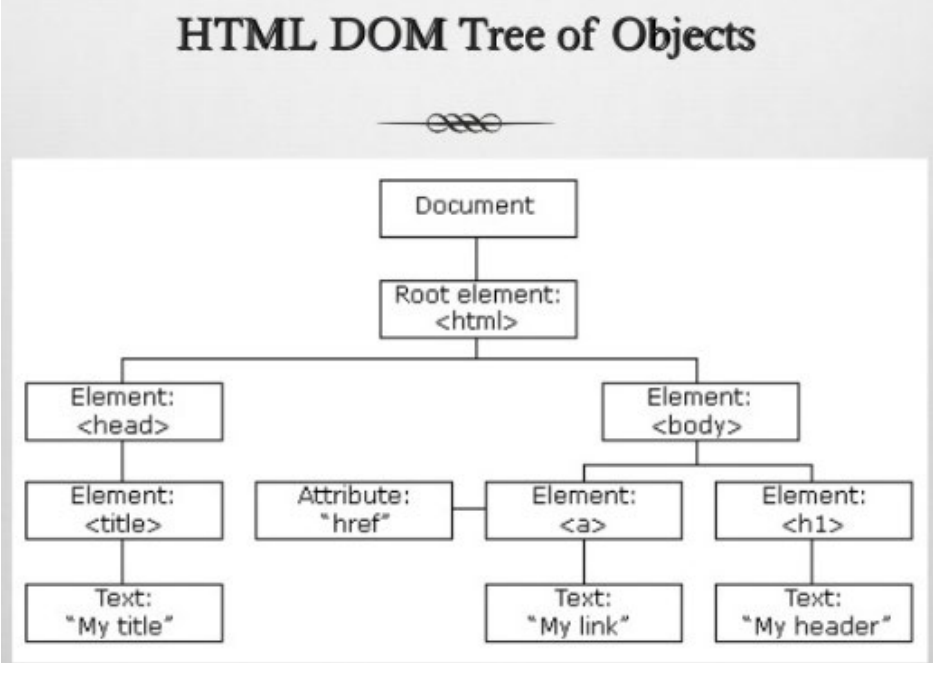
- Beautiful Soup is a python library for pulling data out of HTML and XML files.
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

What is HTML?

- HTML (**H**yper**T**ext **M**arkup **L**anguage) is the standard markup language for document designed to be displayed in a web browser.
- HTML elements are the building blocks of HTML pages. They provide a means to create structured documents.

```
demo.html
1  <!DOCTYPE html>
2  <html>
3  <head>
4  <title>Page Title</title>
5  </head>
6  <body>
7
8  <h1>My First Heading</h1>
9  <p>My first paragraph.</p>
10
11 </body>
12 </html>
```

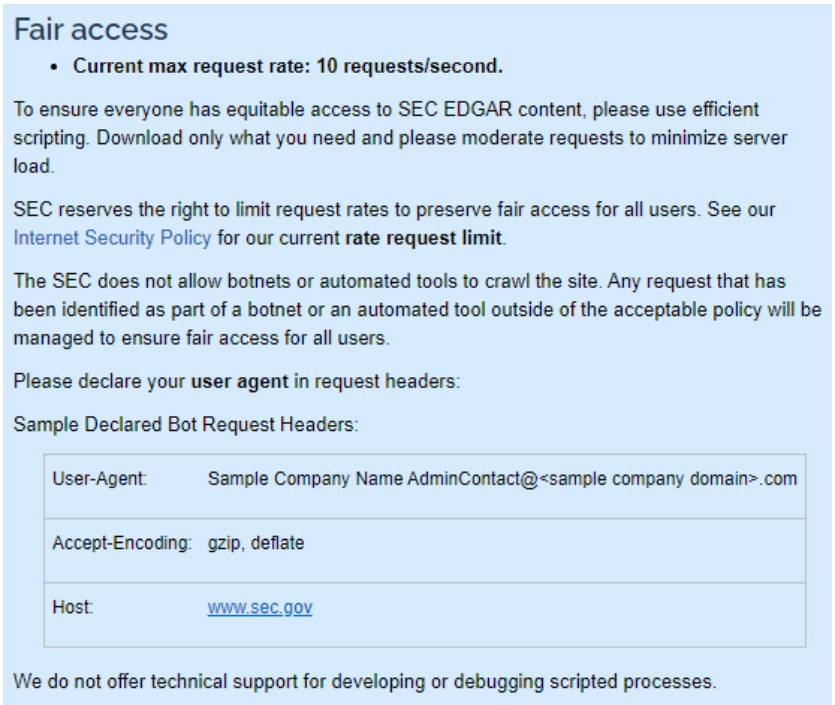




Document Object Model (DOM): a cross-platform and language-independent interface that treats an XML or HTML document as a tree structure wherein each node is an object representing a part of the document

SEC EDGAR

- EDGAR (**E**lectronic **D**ata **G**athering, **A**nalysis and **R**etrieval system) performs automated collection, validation, indexing, acceptance, and forwarding of submission by companies and others who are required by law to file forms with the U.S. SEC (U.S. **S**ecurities and **E**xchange **C**ommission).
- The database contains rich information which is freely available to the public via the internet.



Fair access

- **Current max request rate: 10 requests/second.**

To ensure everyone has equitable access to SEC EDGAR content, please use efficient scripting. Download only what you need and please moderate requests to minimize server load.

SEC reserves the right to limit request rates to preserve fair access for all users. See our Internet Security Policy for our current **rate request limit**.

The SEC does not allow botnets or automated tools to crawl the site. Any request that has been identified as part of a botnet or an automated tool outside of the acceptable policy will be managed to ensure fair access for all users.

Please declare your **user agent** in request headers:

Sample Declared Bot Request Headers:

User-Agent:	Sample Company Name AdminContact@<sample company domain>.com
Accept-Encoding:	gzip, deflate
Host:	www.sec.gov

We do not offer technical support for developing or debugging scripted processes.

More reading: https://www.sec.gov/os/accessing-edgar-data

SEC Filing Types (selected)

- **10-K:**

A 10-K is a comprehensive report filed annually by a publicly traded company about its financial performance in the year and is required by the SEC.

- **10-Q:**

A 10-Q is a comprehensive report filed quarterly by a publicly traded company about its financial performance in the quarter and is required by the SEC.

- **8-K:**

A 8-K report is of unscheduled material events or corporate changes at a company that could be of importance to the shareholders or the SEC. Such as acquisitions, bankruptcy, the resignation of directors, or changes in the fiscal year.

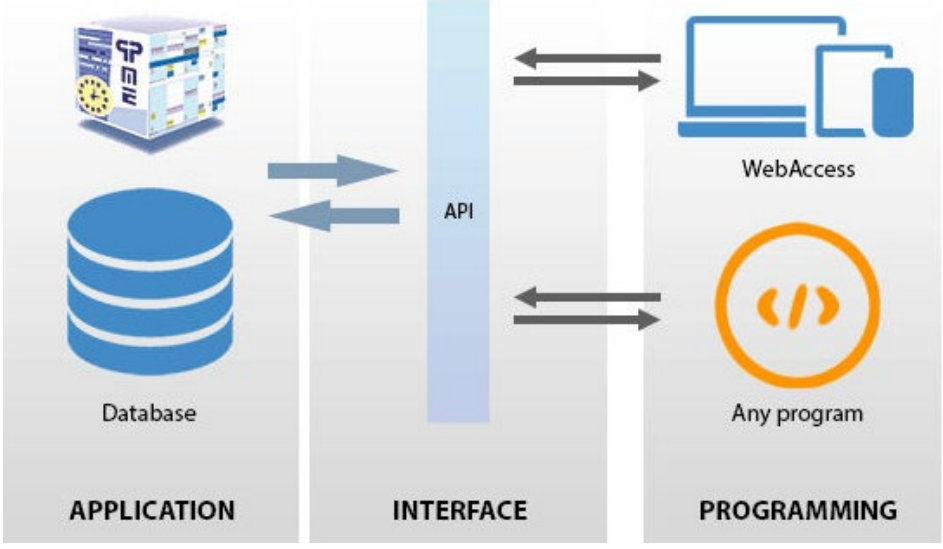
There are many different filings types and can be referenced below.

More reading: https://www.sec.gov/forms, https://en.wikipedia.org/wiki/SEC_filing

API / REST API

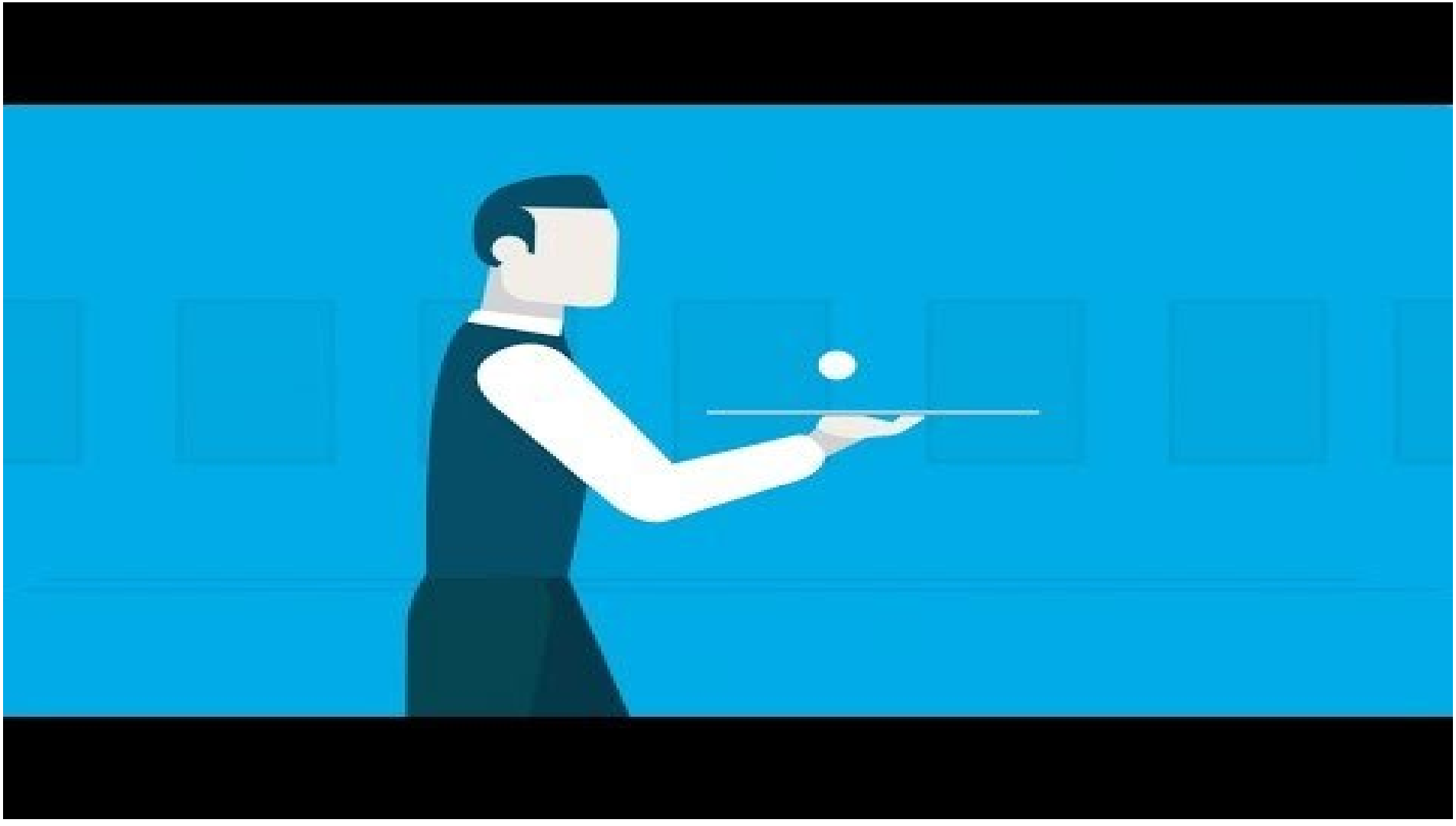
- An **API** is a "*Application Programming Interface*".
- An API is a set of rules and protocols for building and integrating applications.
- Basically an API specifies how software components should interact.
- A good API makes it easier to develop a program by providing all building blocks.
- You can think of an API as a mediator between the users or clients and the resources or web services.
- Usually for organizations to share resources and information while maintaining security, control, and authentication - determining who gets access to what.

HKMA Open API
https://apidocs.hkma.gov.hk/



API / REST API – Cont'd

What is API?



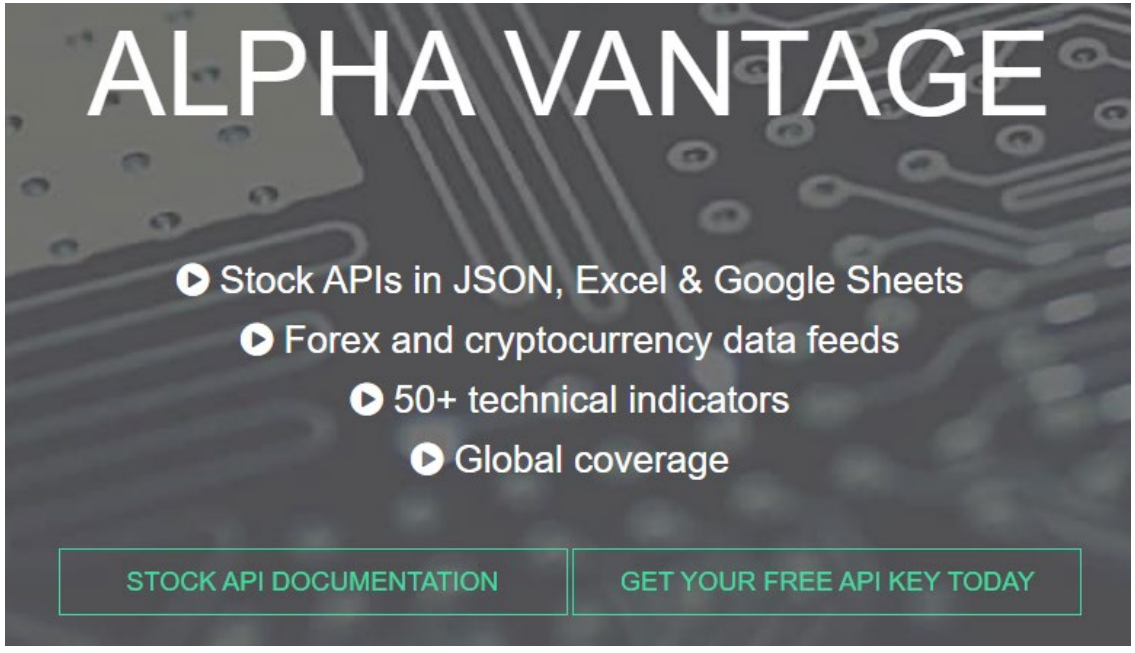
API / REST API – Cont'd

STOCK DATA API

An example:

ALPHA VANTAGE

https://www.alphavantage.co



API / REST API – Cont'd

- REST is **Re**presentational **S**tate **T**ransfer – architectural style for distributed hypermedia systems.
- REST is a set of architectural constraints, not a protocol or a standard. API developers can implement REST in a variety of ways.
- It is a set of rules that developers follow when they create their API.
- It is architectural popular way to structure the backend of a web application.
- In a nutshell, REST determines how the API looks like.
- Web applications are usually divided up into two separate developments team.
  - Front-end – Concentrates on the user experience, the UI, and aspects that people see and interact with.
  - Backend – Server-side development.

An example: InteractiveBrokers

https://www.interactivebrokers.com/en/index.php?f=45185

The End

JUPYTER NOTEBOOK