Data Retrieving for Actuarial and Financial Research

Pasin Marupanthorn

Ph.D. student in Actuarial Mathematics,
Maxwell Institute of Mathematics,
Heriot-Watt University and the University of Edinburgh

Associate Member of Institute of Mathematics and It's Application

Member of Quantitative Risk Solution Laboratory

Member of Strategic Futures Laboratory

OCSC Scholar Student

https://github.com/oporkabbb/TU_Retrieving_Data_Finance_Insurance

Short Bio: Pasin Marupanthorn, AMIMA

Education background:

❑ **B.Sc. Applied Mathematics with Physics,**
   2nd class honor, King Bhumibol Scholarship, Dr. Tab scholarship,
   Thammasat University, TH

❑ **M.Sc. Mathematics**, Dr. Tab scholarship, CEM scholarship, General Research TU funding,
   Thammasat University, TH

❑ **M.Sc. Mathematical Modelling**, Distinction, Outstanding Student Award,
   University of Birmingham, UK

❑ **Ph.D. (Student) Actuarial Mathematics,**
   The Maxwell Institute of Mathematical Science, Heriot-Watt University and University of Edinburgh, UK

Research Areas:

❑ Financial Mathematics – Algorithmic Trading, High Frequency Trading, Portfolio Theory, and Risk Allocation

❑ Actuarial Mathematics – Credibility Theory, and Mortality and Mobility Model

❑ Data Science – Ensemble Learning, Gaussian Process, and Reproducing Kernel Hilbert Space,
   Front - Back End Engineering, Fintech

❑ Simulation – Copula, Cross Entropy Method, and Wang-Landau Sampling

❑ Dynamics System – Continuum Mechanics and Numerical Method
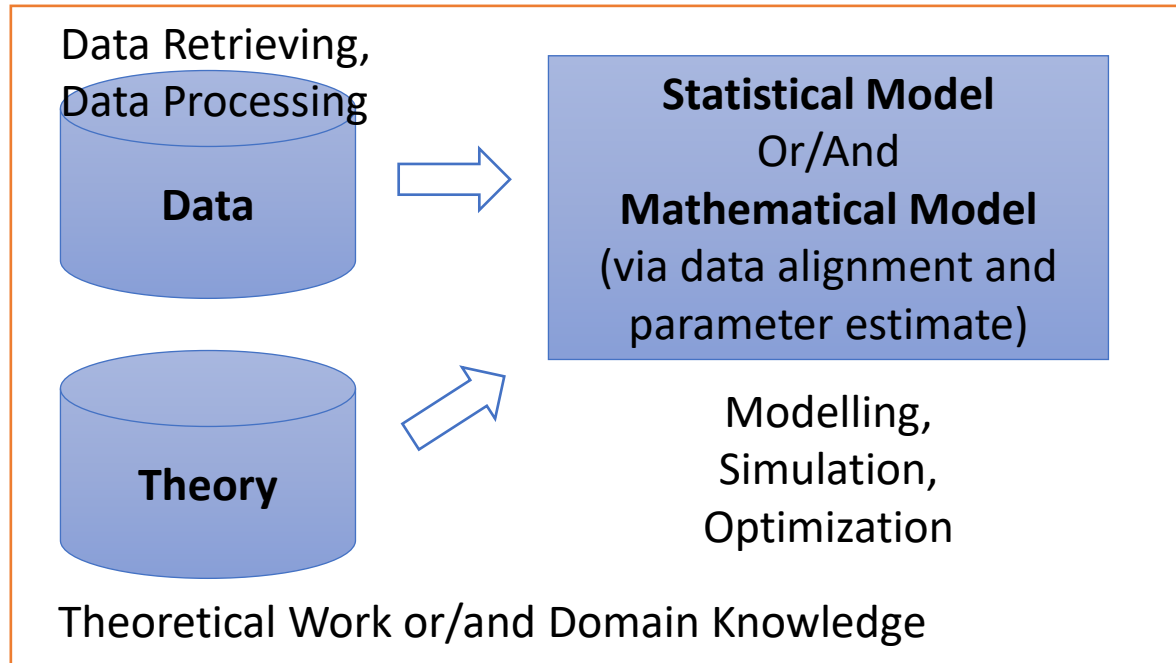
webpage: oporkabbb.wixsite.com/math

Github: https://github.com/oporkabbb

Pasin's Research Flow

Back-End Developing

Front-End Developing

Data Retrieving,
Data Processing

Data

**Statistical Model
Or/And
Mathematical Model
(via data alignment and
parameter estimate)**

Modelling,
Simulation,
Optimization

Theory

Theoretical Work or/and Domain Knowledge

Welcome and Pleasure to Work Collaboratively!!!
Please feel free to contact me at email:
oporkabbb@hotmail.com
Or Facebook:
Pasin Marupanthorn

**Application
or Inventor**

Desktop App, Mobile App,
Package, 3D Printing, IOT

Stakeholders

Distributing,
Github, Web Page,
etc.
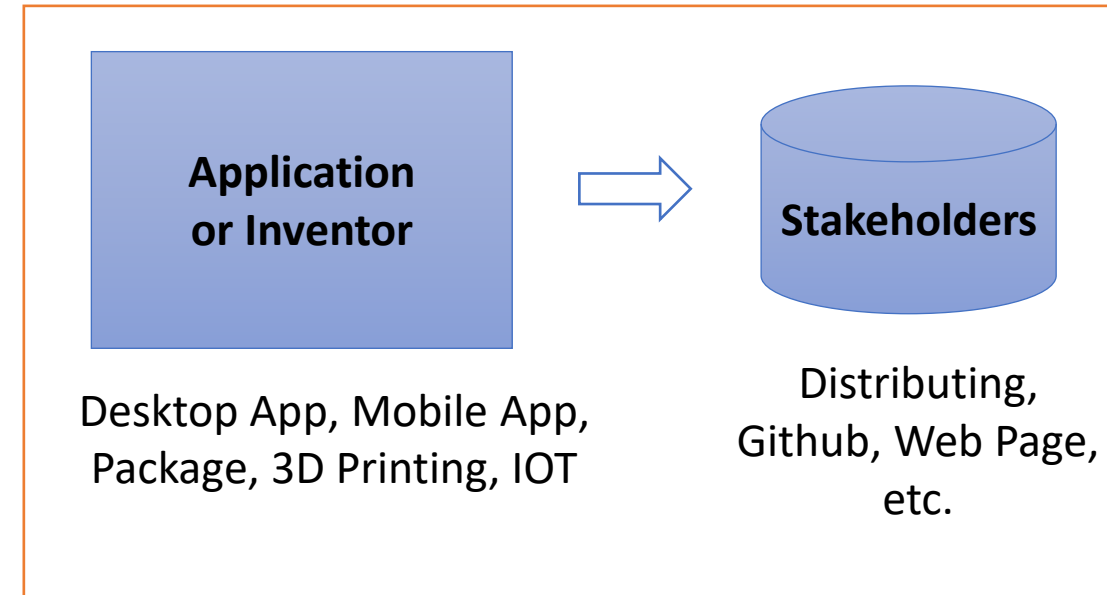
**Warning!! Coding Skill required.
Python is an universal ones.
R is also nice for actuarial and financial works.**

Warning!! How to adapt well in a rapidly changing world

Also see more details in our books

The layout of this page will be finished by ACTA

Machine Learning for Risk and
Insurance: Volume I

Unsupervised Learning Methods

Eric D. Ofosu-Hene,
Gareth W. Peters & Pavel V. Shevchenko
Pasin Marupanthorn

The layout of this page will be finished by ACTA

Machine Learning for Risk and
Insurance: Volume II

Supervised Learning Methods

Eric D. Ofosu-Hene, Gareth W. Peters, Pavel V.
Shevchenko, Pasin Marupanthorn

Outline

- Suggestion Data Resources

- Types of Data Storage Files

- Web Scrapping

- API Connecting

- Case study: Principal Portfolio

1. Suggestion Data Resources

- Financial Data
  - Yahoo Finance: https://finance.yahoo.com/
  - Quandl: https://www.quandl.com/
  - Coingecko Cryptocurrency: https://www.coingecko.com/en
  - World Bank: https://data.worldbank.org/
  - Thai SEC: https://www.sec.or.th/TH/Pages/Home.aspx
  - Thai SET: https://www.set.or.th/th/market/setindexchart.html
  - Thai BOT: https://www.bot.or.th/Thai/Statistics/DataManagementSystem/ReportDoc/DataSetFIFM/Pages/default.aspx

1. Suggestion Data Resources

- Actuarial Data

    - Mortality Rate: https://www.mortality.org/

    - Heath Data: https://www.who.int/data/gho

    - Climate and Environmental Data: https://www.noaa.gov/

    - Google Earth: https://www.google.com/earth/

1. Suggestion Data Resources

- General Data Sets

    - Kaggle: https://www.kaggle.com/

    - Data.World: https://data.world/

    - Our World in Data: https://ourworldindata.org/

    - Thai Open Data: https://data.go.th/

2. Types of Data Storage Files (texts)

- CSV - Comma-Separated Values file. Alternative files XLSX, XLS. Not good for nested data.

```
1
2  id,name,age,gender
3  1,luis,21,m
4  2,john,23,m
5  3,laura,21,f
6  4,beth,25,f
7  5,matt,32,m
```

https://dzone.com/articles/csv-xml-json-in-mulesoft

2. Types of Data Storage Files

- XML - extensible markup language. Used for webpage coding as HTML. Also used for REST API but not popular. Good for nested data.

XML Tree Structure
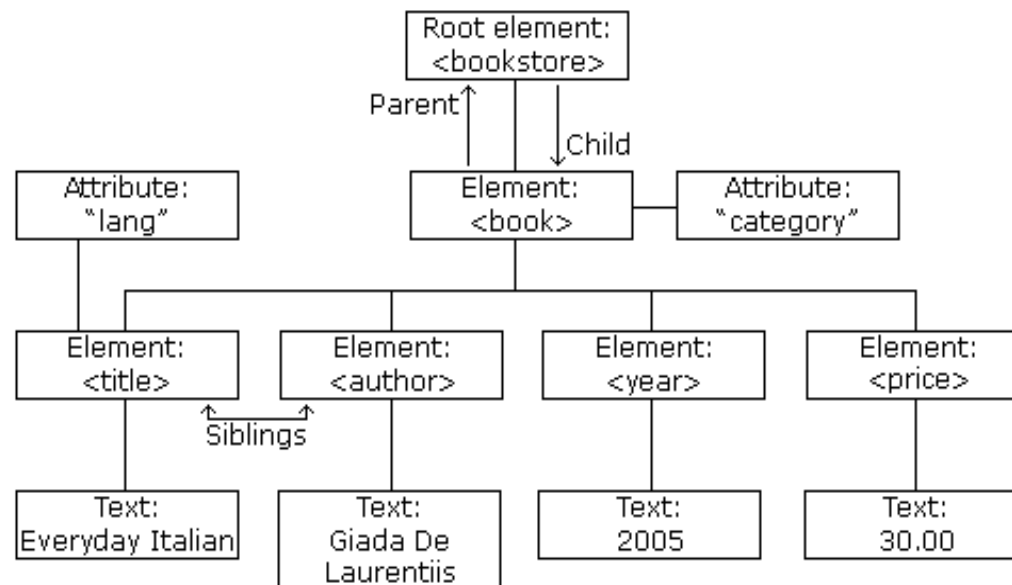
https://www.w3schools.com/xml/xml_tree.asp

```xml
1  <?xml version="1.0"?>
2  <people>
3    <person>
4      <id>1</id>
5      <name>luis</name>
6      <age>21</age>
7      <gender>m</gender>
8    </person>
9    <person>
10     <id>2</id>
11     <name>john</name>
12     <age>23</age>
13     <gender>m</gender>
14   </person>
15   <person>
16     <id>3</id>
17     <name>laura</name>
18     <age>21</age>
19     <gender>f</gender>
20   </person>
21   <person>
22     <id>4</id>
23     <name>beth</name>
24     <age>25</age>
25     <gender>f</gender>
26   </person>
27   <person>
28     <id>5</id>
29     <name>matt</name>
30     <age>32</age>
31     <gender>m</gender>
32   </person>
33 </people>
```

2. Types of Data Storage Files

- JSON- JavaScript Object Notation. Very common used in REST API. Good for nested data.

- SHP - shapefile used in geographic data. Contains shape and location of a geometry.

```
1  {
2     "person": [
3        {
4           "id": "1",
5           "name": "luis",
6           "age": "21",
7           "genre": "m"
8        }
9     ]
10 }
```

2. Types of Data Storage Files

Package in R and python for reading files

File Types	Package in R	Package in python
CSV	util – read.csv() readr – read_csv()	pandas – read_csv()
XLSX, XLS	readxl – read_excel() xlsx - read.xlsx()	pandas - read_excel()
XML, HTML	XML – xmlParse() xml2 - read_xml() rvest - read_html()	xml.dom - minidom.parse() codecs - open() bs4 – BeautifulSoup()
JSON	jsonlite - read_json() rjson – fromJSON()	json – loads()
SHP	sf – st_read() rgdal - readOGR()	fiona – open() pyshp - shapefile.Reader()

2. Types of Data Storage Files

Using Thai Language in R and python

R	python
Sys.setlocale("LC_CTYPE", "thai") options(encoding="UTF-8")	#-*-coding: utf-8 -*-

Let try!!

https://data.go.th/dataset/university
See HowtoReadFile.R
- How to read CSV, XML and JSON
files by R

3. Web Scraping

Suggest Scrapping Packages

R	Python
rvest htmltab	bs4 or BeautifulSoup

Web Scraping is the process of collecting structured **web** data in an automated fashion. It's also called **web** data extraction.

https://www.zyte.com/learn/what-is-web-scraping/#:~:text=Web%20scraping%20is%20the%20process,market%20research%20among%20many%20others.

HTML Structure similar to XML

https://designshack.net/articles/html/what-is-html-the-anatomy-of-an-html5-document/
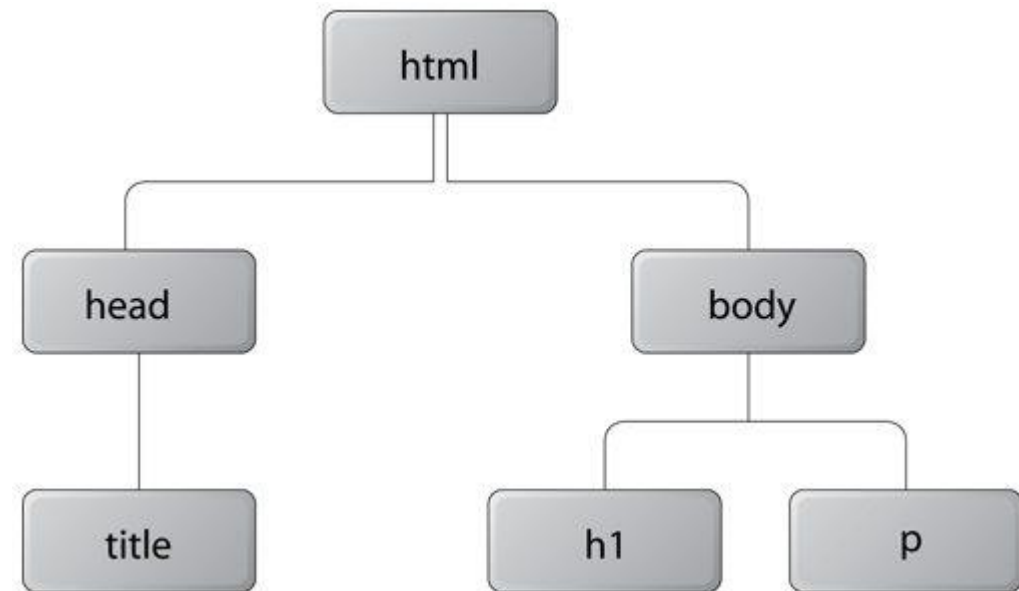
3. Web Scraping

In Google Chrome, right-click and go to Inspect

3. Web Scraping

In Google Chrome, right-hand window is a HTML structure of this page.
- <a> is a node
- href is an attribute
We have many ways to point out the location of the data we want such as node, attribute or xpath – specific location.
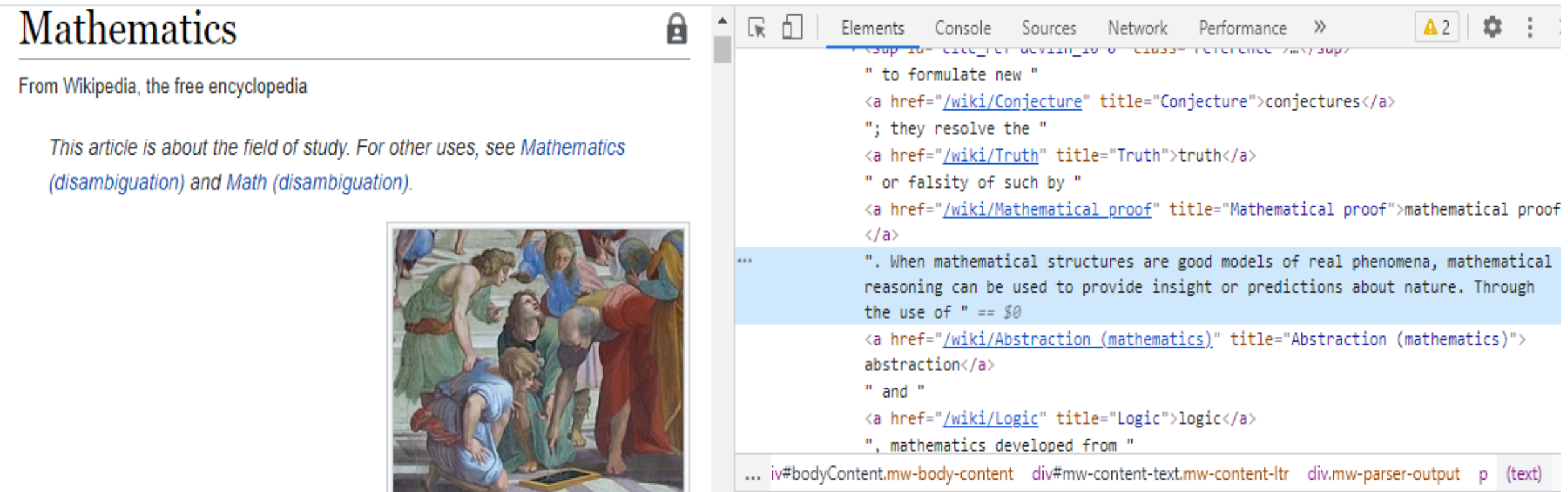
Let try!!

https://en.wikipedia.org/wiki/Life_expectancy
See WiKiScrap.R
- How to get data by Web scrapping

https://www.set.or.th/set/commonslookup.do?language=en&country=US
See SETscrap.R
- Example for real usage scarping

Note!! One of disadvantage of web scraping is that the HTML structure of web page may change by developer. We need to update our code regularly.

4. API connecting

API - Application Programming Interface

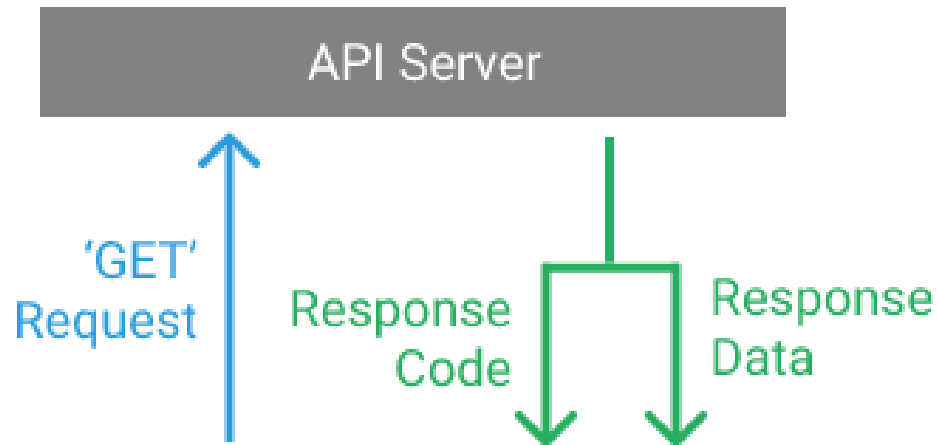| R packages	python packages
"httr" - GET	"requests" - get
"jsonlite"	"json"

'GET' Request | Response Code | Response Data

API Server

https://www.dataquest.io/blog/python-api-tutorial/
https://www.dataquest.io/blog/r-api-tutorial/

- Request forms are different depended on provider. Most of forms are in JSON format
- Similar, response data is usually JSON format
- Example of Request forms:

```
#Candlestick Data
data <- GET(
  url = "https://api.binance.com",
  path = "/api/v3/klines",
  query = list (
    symbol = "LTCBTC",
    interval ="1m",
    limit = "60"
  )
)
```

Some Methods to access data via API
1. Accessing URL as JSON
2. Using GET method
3. Using provided package

4. API connecting

List of some API providers. Many Thai gov organizations have provided API.

Name	Site
NOAA	https://www.ncdc.noaa.gov/cdo-web/webservices/v2
Quandl	https://www.quandl.com/tools/api
Twister	https://developer.twist.com/v3/
Binance	https://binance-docs.github.io/apidocs/spot/en/#change-log
Thai Open Data	https://opend.data.go.th/register_api/
Thailand Meteorological Department	https://data.tmd.go.th/api/index1.php
Department of Disease Control	https://covid19.th-stat.com/th/api
Ministry of Commerce Thailand	https://data.moc.go.th/
Settade	https://developer.settrade.com/open-api/

4. API connecting

More convenient!! Some API providers or developers have a package to connect to their API

Name	R packages	python packages
NOAA	rnoaa	noaa-sdk
Quandl	quandl	quandl
Twister	twitteR	tornado-swagger
Binance	binancer	python-binance
Settrade	-	settrade

Let try!!

https://covid19.th-stat.com/th/api
https://data.tmd.go.th/api/index1.php

See ThaiAPI.R
-    To get update Covid cases in Thailand via API
-    To get climate data from Thailand Meteorological Department

https://binance-docs.github.io/apidocs/spot/en/#kline-candlestick-data
See BinanceAPI.R
- To get candlestick Data of bitcoin

Other Methods to Get Online Data

- WebDriver

- WebSocket API

- Etc.

5. Case Study: Principle Portfolio

- **Step 1:** Retrieve stock data in the index FTSE100 ETF from Yahoo Finance

- **Step 2:**

- Apply PCA – Principal Component Analysis to the Data (Covariance Matrix)

- Apply kPCA – Kernel Principal Component Analysis to the Data (Non-Linear Covariance Matrix with RBF kernel)

- **Step 3:** Use normalized eigenvector as the investment weight for each asset in portfolio

$$\boxed{\text{Port}} = \lambda_1 \boxed{\text{Stock 1}} + \lambda_2 \boxed{\text{Stock 2}} + \lambda_3 \boxed{\text{Stock 3}} + ... + + \lambda_{100} \boxed{\text{Stock 100}}$$

See PrinciplePortfolio.R