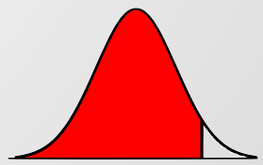
Machine Learning & FinTech
Python and EDA

Huei-Wen Teng

Department of Information Management and Finance
National Yang Ming Chiao Tung University
https://hackmd.io/@hwteng/HyKOPoA6d

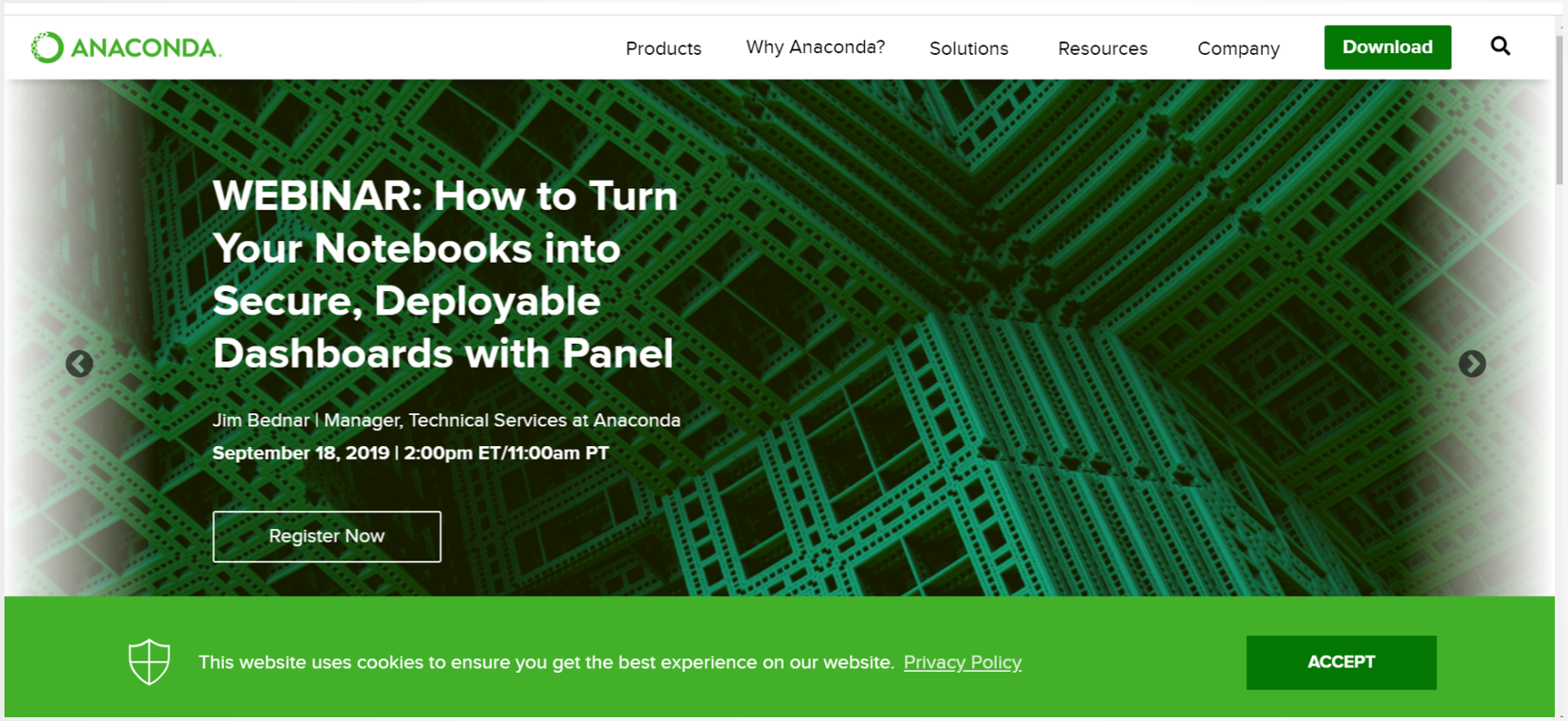
Outline

1. Python
2. EDA
3. More



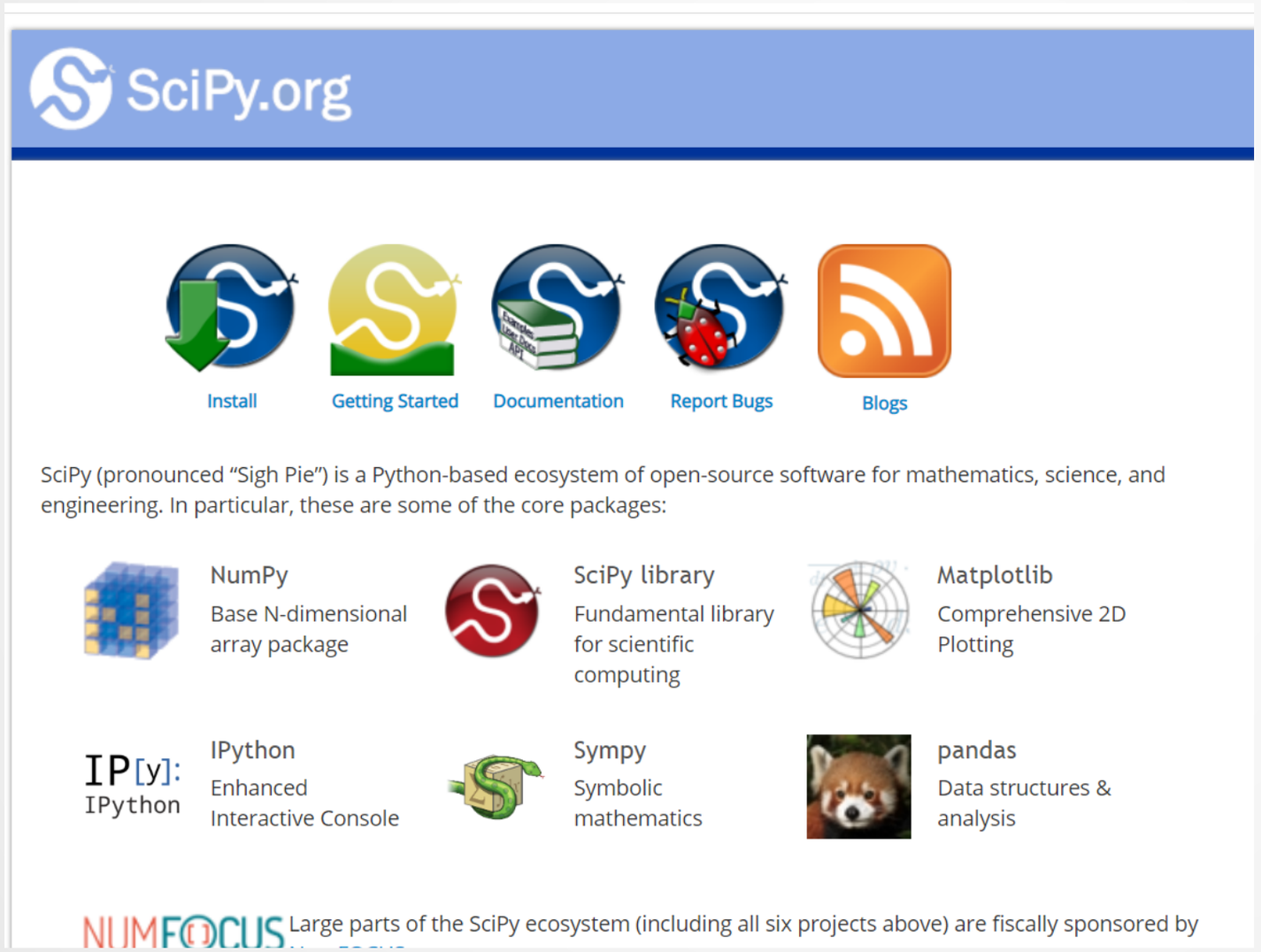
Motivation

- Download Anaconda
  - Spyder: similar to Matlab
  - jupyter (used for course demonstrations)
  - Colab (google, the simplest and slowest)



SciPy

- SciPy (pronounced “Sigh Pie”) is a Python-based ecosystem of open-source software for mathematics, science, and engineering.

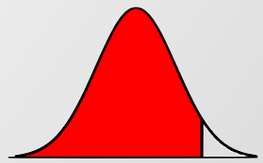


- Reading: Scipy Lecture Notes

NymPy

- NumPy is the fundamental package for scientific computing with Python. It contains among other things:
- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

- Reading: Numpy Quick Start Tutorials



Pandas

- Pandas stands for “Python Data Analysis Library”.
  According to the Wikipedia page on Pandas, “the name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.”
- Pandas is for data munging
- It proides Series and DataFrame data strucutre

matplotlib

- Matplotlib is a Python 2D plotting library.

matplotlib
Version 3.1.1

home | examples | tutorials | API | contents »

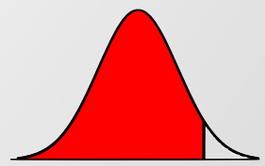
Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the `pyplot` module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

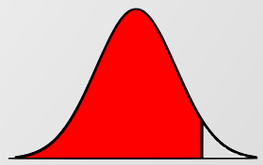
Installation

1_9IU5fBzJisilYiR....png ^ | 1_5Uza5wbRm....jpeg ^ | scipy-lectures-sci....zip ^



Reading: Python basics

- #python_1_basics.ipynb
- #python_2_numPy.ipynb
- #python_3_pandas.ipynb



Exploratory Data Analysis (EDA)

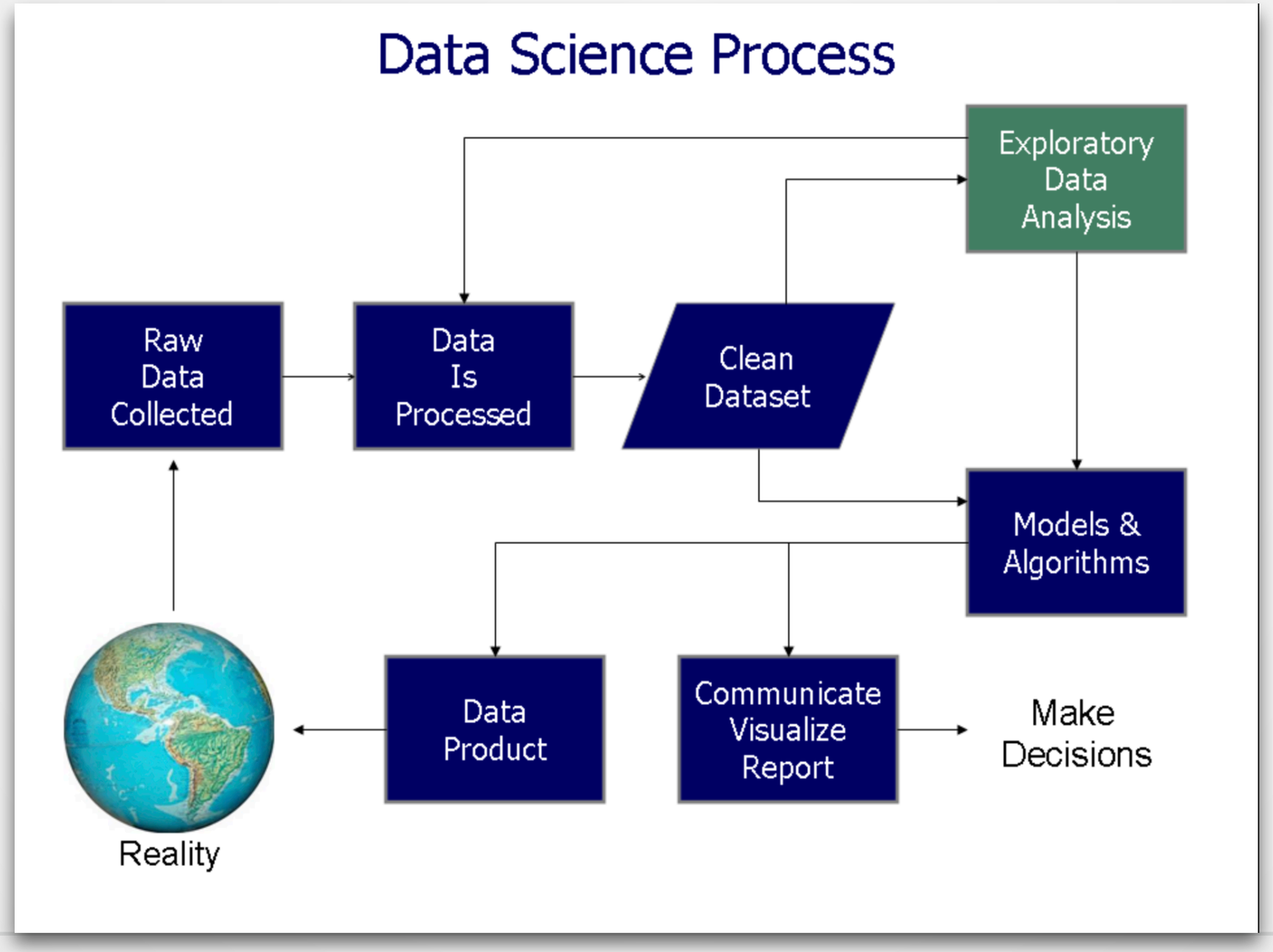
- EDA was promoted by John Tukey (FFT and box plot) to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.
- Equal to
  - Data visualizaion
  - Data mining
  - Unserpervised learning

https://en.wikipedia.org/wiki/John_Tukey

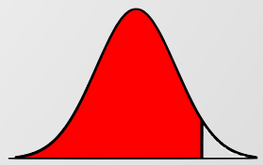
A nutshell

- The EDA tools allow researchers to obtain a general impression of their data.
  - Recall Chapters 1 to 3 in Statistics 101
  - Simple **numerical presentation**: mean, var, std, max, min, quantiles.
  - Simple **graphical presentation**: histogram, boxplots, scatterplots.
- What's more?
  - **Sophisticaed numerical presentation**: statistical tests (ADF test, LM test)
  - **Sophisicated Graphical presentation**: dimensinoal reduction (PCA), clustering, etc.

Data Science Process (This course!!!)
Job market for Data science is dying!
But job market for business analystics is booming!!!



https://en.wikipedia.org/wiki/Exploratory_data_analysis



The objectives of EDA

- The objectives
  - Enable unexpected discoveries in the data
  - Suggest hypotheses about the causes of observed phenomena
  - Assess assumptions on which statistical inference will be based
  - Support the selection of appropriate statistical tools and techniques
  - Provide a basis for further data collection through surveys or experiments
- Also known as data mining or data exploration!

Example

- This research aimed at the case of customer's default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.
- Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.
- **#EDA_credit_default.ipynb**

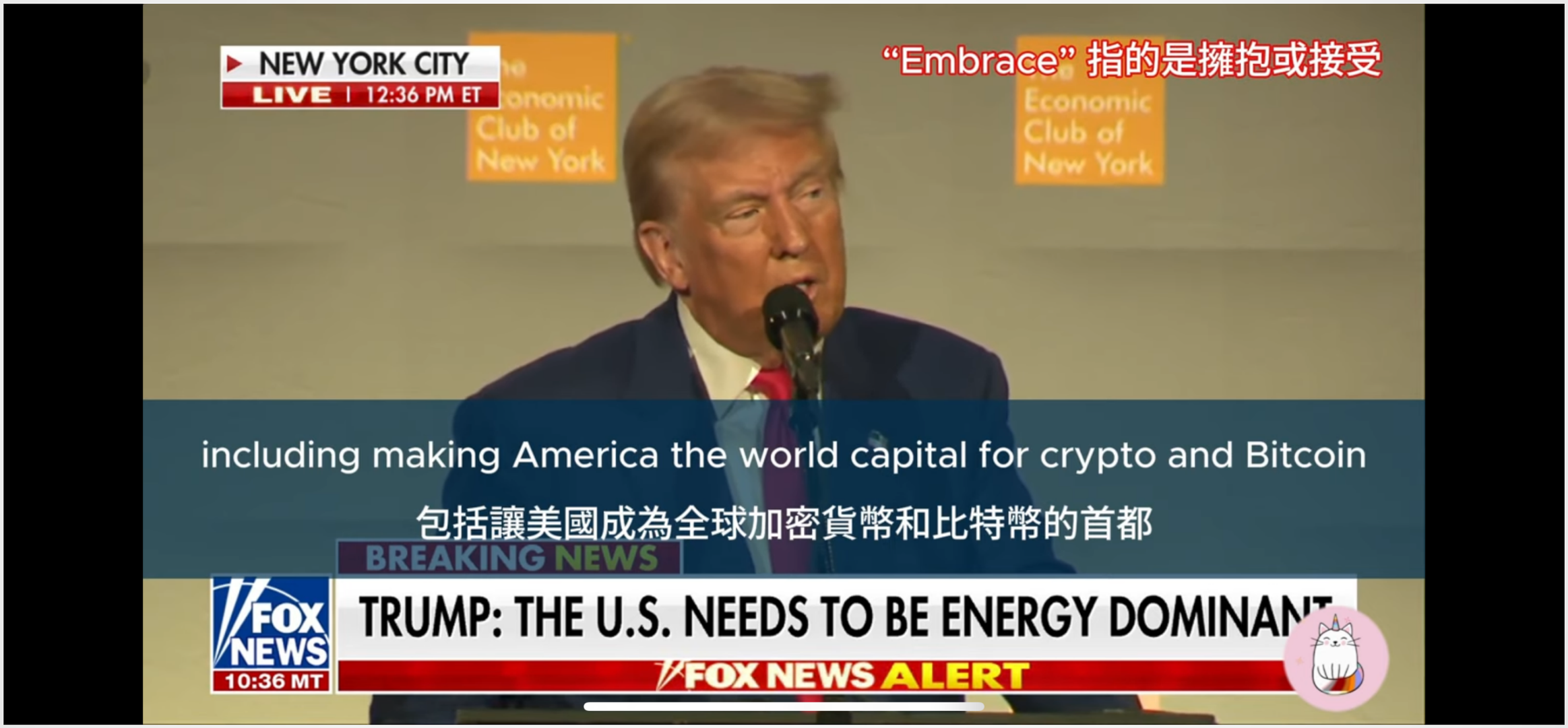
More examples on Colab

- #EDA on GMC https://colab.research.google.com/drive/1a6RNuvZnxdPrsGqXX9EOJuzC0D8QEuhs?usp=sharing

- #EDA on TW stocks https://colab.research.google.com/drive/1EMSLOb0TaaLAiDUN80G4MBWHGoK4sbpQ?usp=sharing

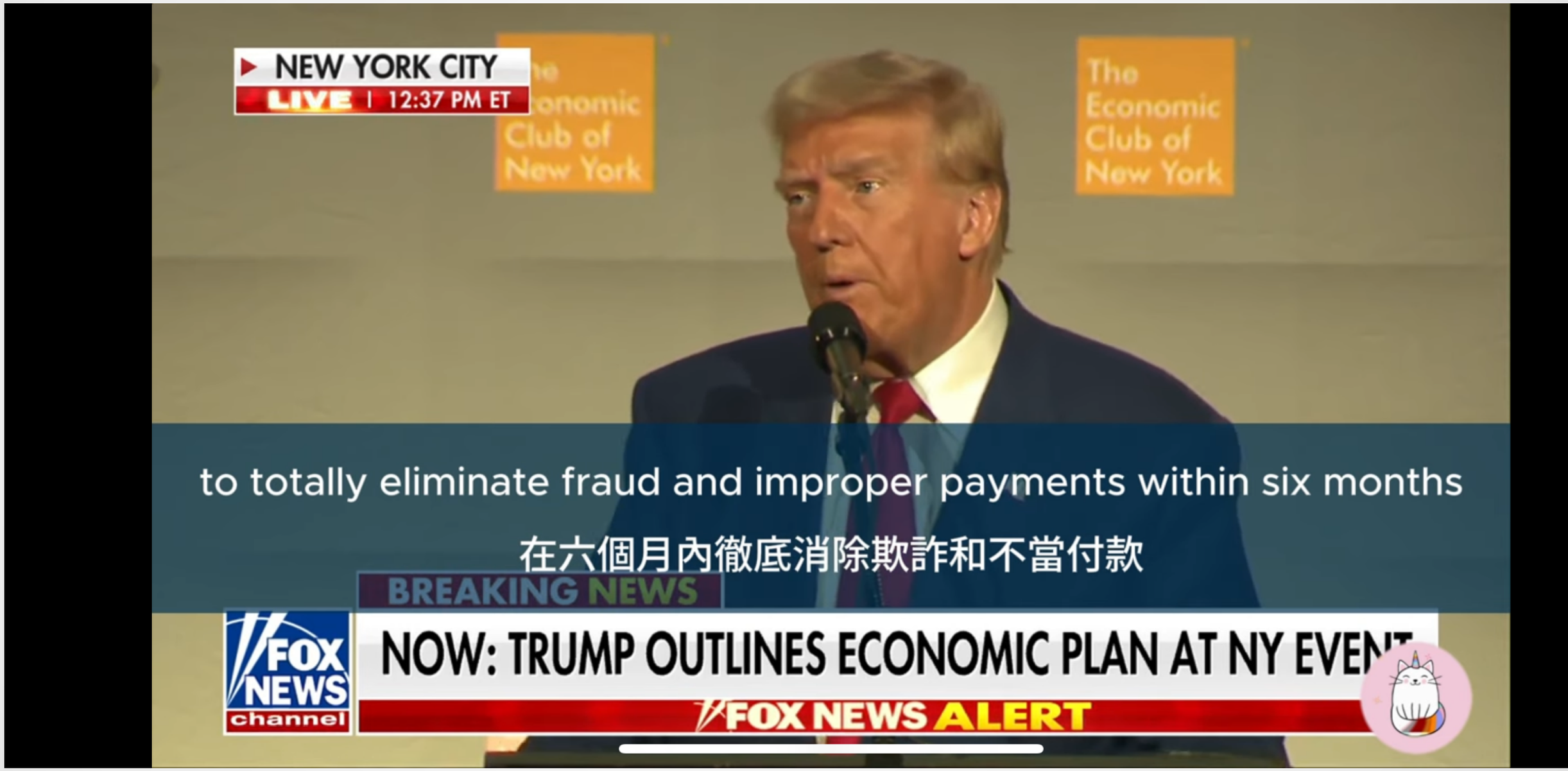
Policies and regulations

20240906



Policies and regulations

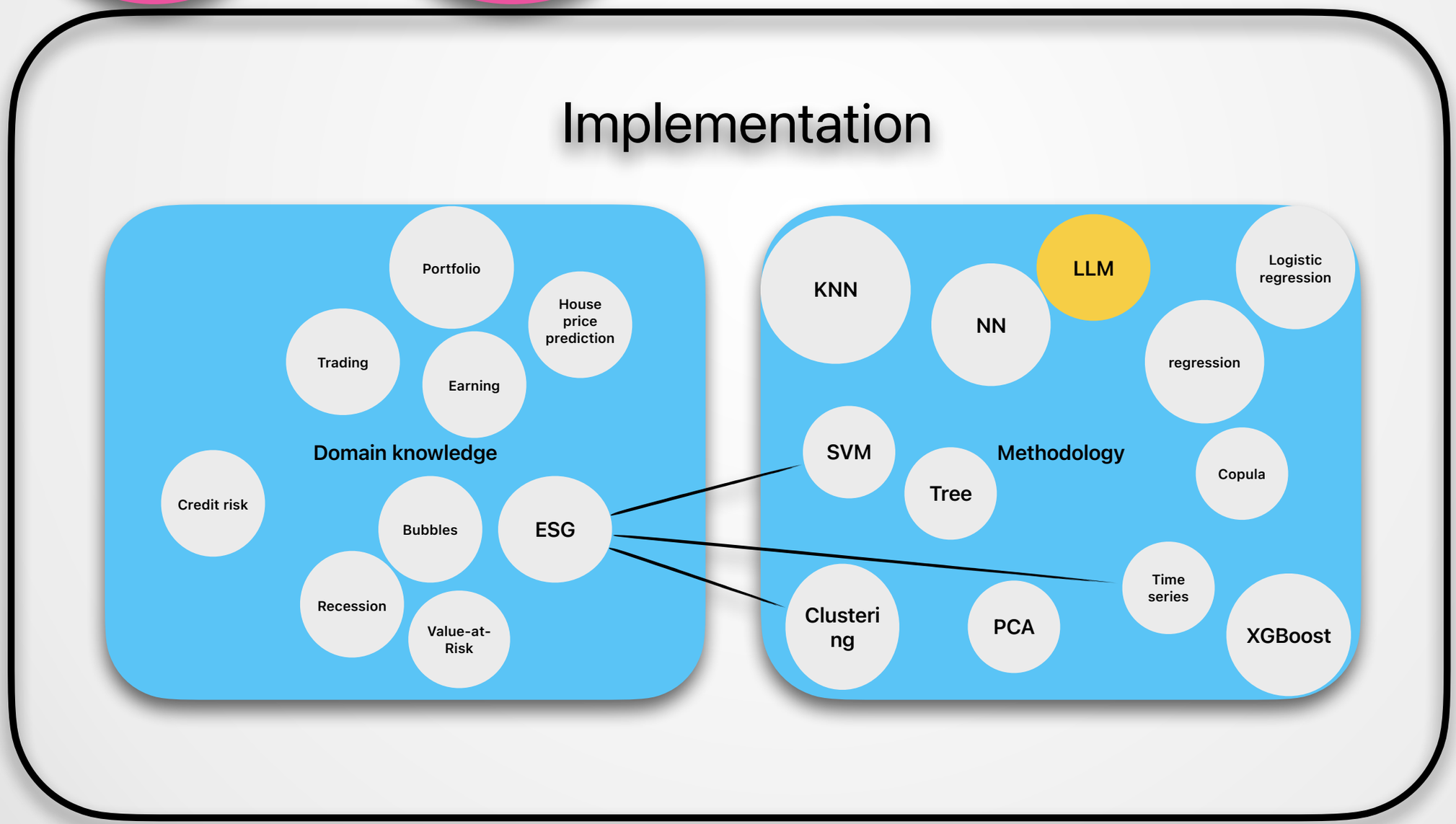
20240909



Implementations

Passions & Curiosity

Regulations



Data sets (will be downloaded soon in GitHUB)

- Credit scoring
  - Mabel's first dataset
  - Mabel's second dataset
  - Pauls' four datasets
- Fraud detection
  - Crissy's dataset
- BTC/cryptocurrencies prediction
  - Yenting's dataset: BTC and other exploratory variables
  - Wendy's dataset: Cryptocurrencies market
  - Jason's dataset: Taiwan stock market
- If you are not interested in the above, chat with me before hand. With convincing and promising reasons, you can use your own.

Other source of data

- Kaggle
  - IEEE-CIS Fraud Detection
  - House Prices: Advanced Regression Techniques
- UCI Machine Learning Repository
  - Bank Marketing Dataset
  - Default of credit card clients dataset

NYCU paid dataset:

- TEJ

- WRDS

Machine Learning & FinTech
Python and EDA

Huei-Wen Teng

Department of Information Management and Finance
National Yang Ming Chiao Tung University
https://hackmd.io/@hwteng/HyKOPoA6d