



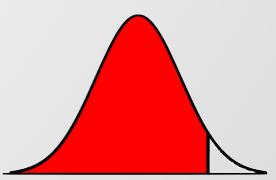
Machine Learning & FinTech Python and EDA

Huei-Wen Teng

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

Outline

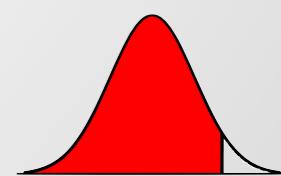
1. Python
2. EDA
3. More



Motivation

- Download Anaconda
 - Spyder: similar to Matlab
 - jupyter (used for course demonstrations)

The screenshot shows the official website for Anaconda. At the top, there's a navigation bar with links for Products, Why Anaconda?, Solutions, Resources, Company, a prominent green "Download" button, and a search icon. The main content area has a dark green background with a complex, fractal-like pattern. In the center, there's a large white box containing text about a webinar. The text reads: "WEBINAR: How to Turn Your Notebooks into Secure, Deployable Dashboards with Panel". Below this, it says "Jim Bednar | Manager, Technical Services at Anaconda" and "September 18, 2019 | 2:00pm ET/11:00am PT". At the bottom of this box is a "Register Now" button. At the very bottom of the page, there's a green footer bar with a shield icon and the text: "This website uses cookies to ensure you get the best experience on our website. [Privacy Policy](#)". To the right of this text is a "ACCEPT" button.

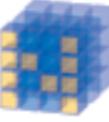


SciPy

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

The screenshot shows the SciPy.org homepage. At the top is a blue header bar with the SciPy logo and the text "SciPy.org". Below the header are five navigation icons: "Install" (blue circle with a green arrow), "Getting Started" (yellow circle with a green and yellow "S"), "Documentation" (blue circle with a white book and "S"), "Report Bugs" (blue circle with a red bug and "S"), and "Blogs" (orange square with an orange RSS icon). Below these icons is a section titled "SciPy (pronounced ‘Sigh Pie’) is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:" followed by six entries: NumPy (base N-dimensional array package), SciPy library (fundamental library for scientific computing), Matplotlib (comprehensive 2D plotting), IPython (enhanced interactive console), Sympy (symbolic mathematics), and pandas (data structures & analysis). A "NIJMFOCUS" logo is at the bottom left, and a note states "Large parts of the SciPy ecosystem (including all six projects above) are fiscally sponsored by".

SciPy (pronounced “Sigh Pie”) is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

 NumPy Base N-dimensional array package	 SciPy library Fundamental library for scientific computing	 Matplotlib Comprehensive 2D Plotting
 IP[y]: IPython Enhanced Interactive Console	 Sympy Symbolic mathematics	 pandas Data structures & analysis

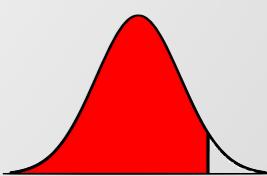
NIJMFOCUS Large parts of the SciPy ecosystem (including all six projects above) are fiscally sponsored by

- Reading: Scipy Lecture Notes



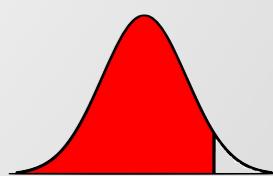
NymPy

- 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

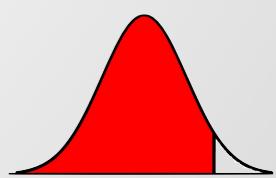
- Pandas stands for “Python Data Analysis Library”.
According to the Wikipedia page on Pandas, “the name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.”
- Pandas is for data munging
- It provides Series and DataFrame data structure
- It helps to retrieve different formats of data



matplotlib

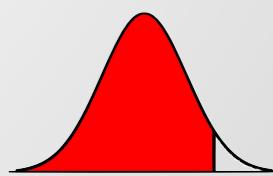
- Matplotlib is a Python 2D plotting library.

The screenshot shows the official Matplotlib website. At the top is the large blue "matplotlib" logo with a circular icon containing colored bars. Below it is the text "Version 3.1.1". A horizontal navigation bar contains links for "home", "examples", "tutorials", "API", and "contents". The main content area starts with a paragraph about Matplotlib's capabilities: "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." Below this are four small square thumbnails of plots: a line plot with multiple oscillations, a histogram with a normal distribution curve, a heatmap with a central peak, and a 3D surface plot. Further down, there is a section titled "Installation" with three download links: "1 9IU5fBzJisilYiR....png", "1 5Uza5wbRm....ipea", and "scipy-lectures-sci....zip".



Reading: Python basics

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

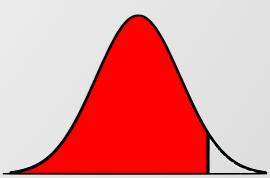


EDA

- Exploratory data analysis 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.

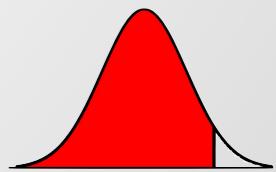


https://en.wikipedia.org/wiki/John_Tukey

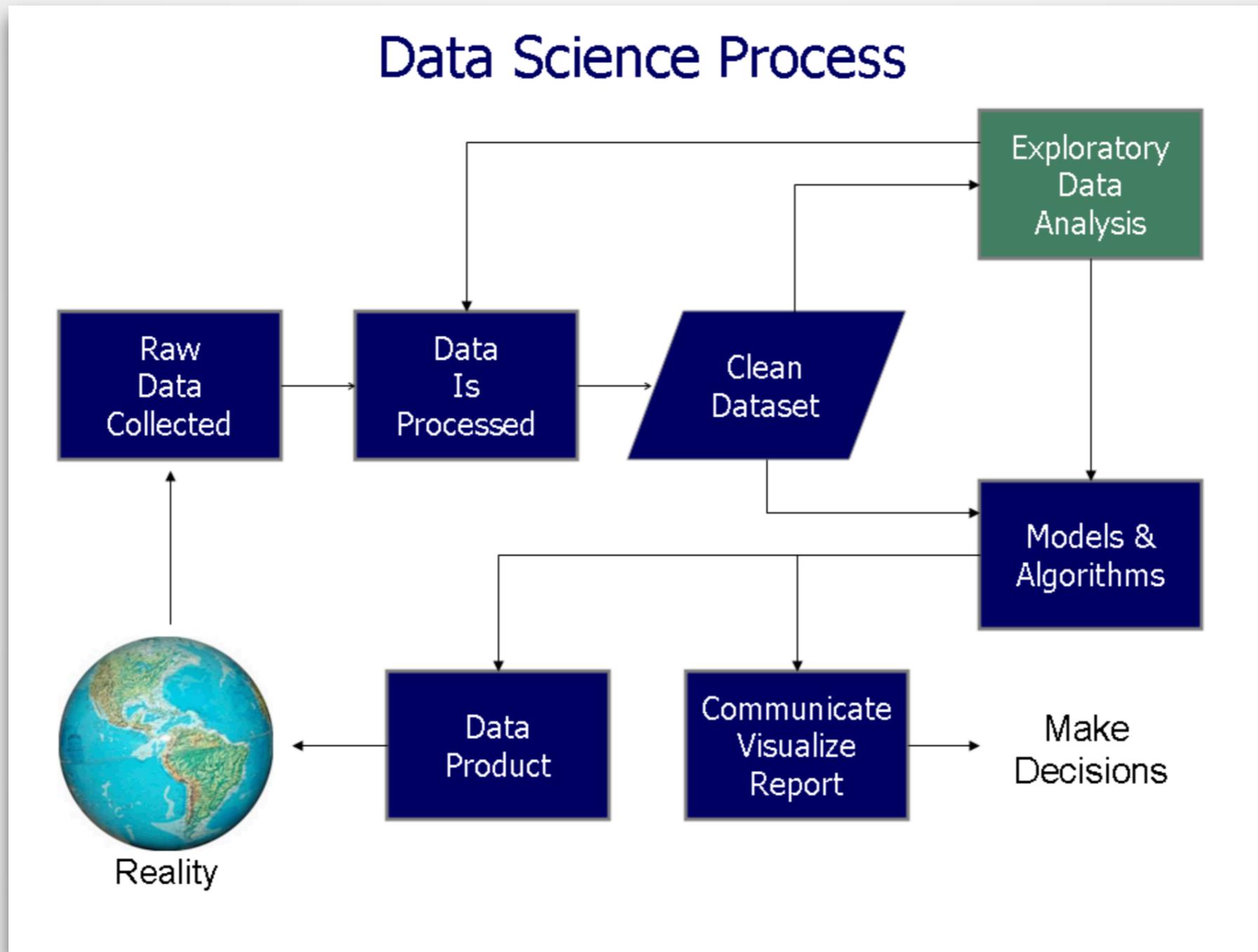


A nutshell

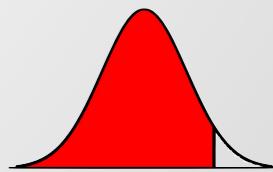
- The EDA tools allow researchers to obtain a general impression of their data.
 - Recall Chapters 1 to 3 in Statistics 101
 - Simple numerical presentation: mean, var, std, max, min, quantiles.
 - Simple graphical presentation: histogram, boxplots, scatterplots.
- What's more?
 - Sophisticated numerical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



Data Science Process

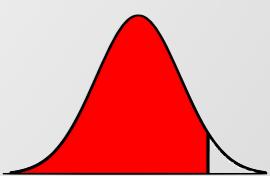


https://en.wikipedia.org/wiki/Exploratory_data_analysis



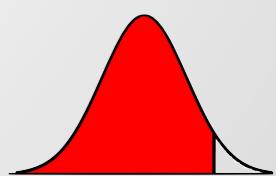
The objectives of EDA

- 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!**



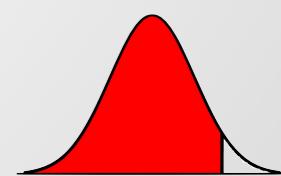
Example

- 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. Data available at <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>.
- See EDA_credit_default.ipynb



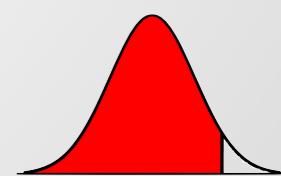
Policies and regulations

20240906



Policies and regulations

20240909

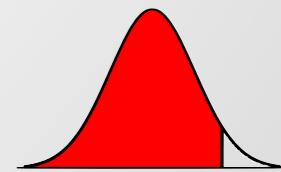
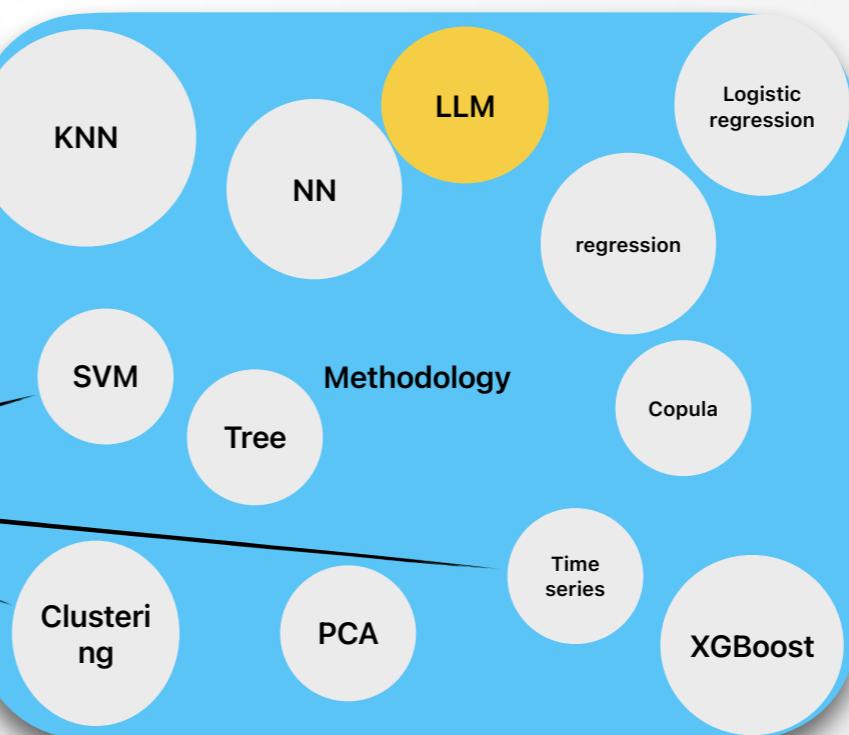
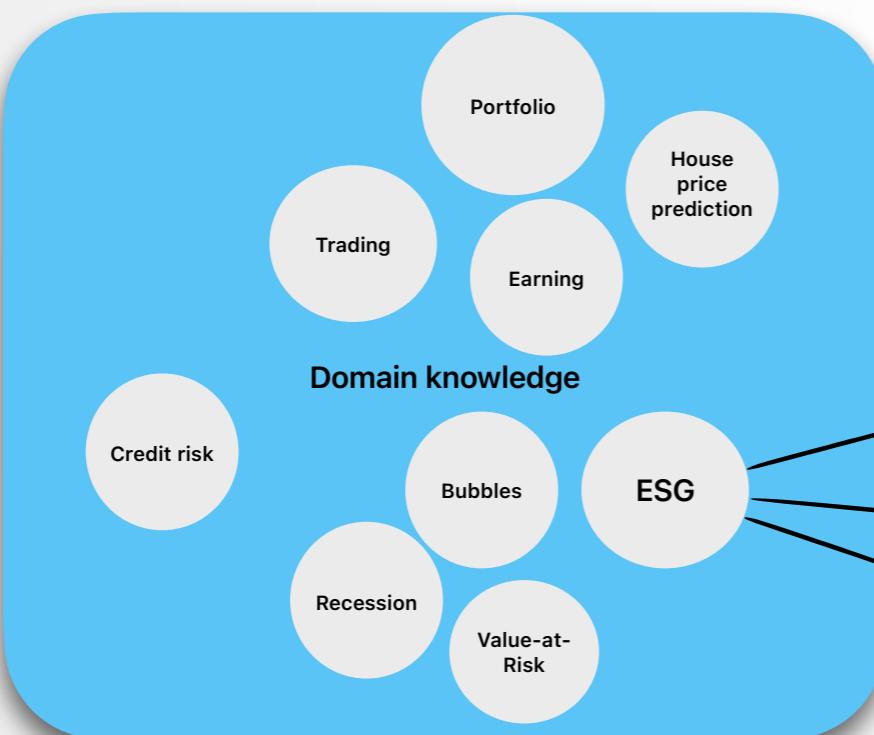


Implementations

Passions & Curiosity

Regulations

Implementation





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>