

AI & ML for Data Scientists

Class 4: **Real** Big Financial Data

Ray Ge

Quant RUC (人大量化) & IE Finance

January 30, 2026



Real Financial Data
ooo

WRDS
ooooo

Kaggle
oooo

Selenium
oooooo

Real Financial Data

WRDS

Kaggle

Selenium

Real Financial Data

WRDS

Kaggle

Selenium

Real Financial Data, Why?

- What is the most important elements for Machine Learning?
Data
- What makes the ML in finance unique? (we financial data)
- Why real data?

First look at the fake data

- sklearn.datasets is a good source for TOY data
- Good source for practice
- Only issue is that fake data is fake
- Lets check out why (Please follow to blank Jupyter)

Real Financial Data
ooo

WRDS
●oooo

Kaggle
oooo

Selenium
oooooo

Real Financial Data

WRDS

Kaggle

Selenium

WRDS

The screenshot shows the homepage of the Wharton Research Data Services (WRDS) website. The header features the Wharton University of Pennsylvania logo, the WRDS acronym, and the full name "WHARTON RESEARCH DATA SERVICES". A search bar with a magnifying glass icon and a dropdown menu are also visible. Below the header, a navigation bar includes links for "Get Data", "Analytics", "Classroom", "Videos", and "Research". On the left side, there is a sidebar titled "VENDORS" listing "CRSP", "Compustat - Capital IQ", "IBES", and "TAQ". On the right side, there is a sidebar titled "DATA DICTIONARY & UPDATES" listing "Find Companies", "Find Variables", and "Linking Databases". The main content area is currently empty.

答:

WRDS

- professional level financial data for stock & company study
- used by both financial companies and financial researchers

WRDS: easy to use

- Easy to use especially for Python users
- We can use both UI and API (what is UA and API?)
- its check it with me step by step and login from lib

WRDS

Please follow me to use WRDS

Real Financial Data
ooo

WRDS
ooooo

Kaggle
●ooo

Selenium
oooooo

Real Financial Data

WRDS

Kaggle

Selenium

Kaggle

- Kaggle, a subsidiary of Google LLC
- Heavily platform for Quant Research (us)
- Codes, data, competition and more
- Let check it out! (Kaggle)

Kaggle

- Kaggle is most important data source for now
- You can search and find your interested research topics
- Let check it out! (Kaggle)

Huggingface

The screenshot shows the Hugging Face homepage. At the top left is the user profile icon and the text "Hugging Face". A search bar contains the placeholder "Search models, datasets, users...". To the right are links for "Models" and "Datasets". On the far right are standard browser navigation icons.

The main area has a sidebar on the left with the following sections:

- New** (button)
- jedibear**
 - Profile
 - Inbox (0)
 - Settings
 - Billing
 - Get PRO
- Organizations**
 - Create New
- Resources**

The main content area displays the user's "Following" list, which includes 6 items:

- Following 6**
 - All Models Datasets Spaces Papers Collections
 - Community Posts Upvotes Likes Articles
- nvidia model updated by Victor49152** about 3 hours ago
 - nvidia/Qwen3-VL-235B-A22B-Instruct-NVFP4-M...**
119B · Updated 2 days ago · ↴ 39
- FranckAbgrall published a changelog** about 6 hours ago
 - Changelog**
 - View Running Jobs Count from the User Menu

Real Financial Data
ooo

WRDS
ooooo

Kaggle
oooo

Selenium
●ooooo

Real Financial Data

WRDS

Kaggle

Selenium

Data from the internet

1. Internet has valuable data for the financial predictions
2. Internet data low quality? No
3. Selenium is a powerful and popular tool

But how to use?

- I will guide you to study this package
- but next time you should know how to learn any package by yourself

But how to use?

- Template + Documentation + CHATGPT (Gemini) + Google
- Template (from search bing and from CSDN, StackOverFlow, CHATGPT)
- Unknown knowledge → Bing + Documentation + ChatGPT

Please follow me to the selenium codes

CH1_Class4_Tesla_Earning_Call.ipynb

HW3: Practice Selenium

- How to practice Python → Selenium
- Find a interesting source
- Get data by using selenium
- Do data description
- OLS or other model to showcase your research idea