# Financial Data Analysis with Python

## Lecture 01. Introduction

Luping Yu (俞路平)

Xiamen University

February 22, 2022

# Course Information

- Instructor: Luping Yu (俞路平)
  - B.Eng. Electronic Information Engineering, UESTC
  - M.Sc. Computer Science, Bristol
  - Ph.D. Finance, HKU
- Email: lupingyu@xmu.edu.cn
- Office: J2-326

- DingTalk Group: 32027301
- Tutor: Wenjie Lu (卢文杰)
- Any issues on administration (e.g., enrollment, time clash, lab entrance, absence from the exams, etc.) and homework (e.g., clarification of problems) should contact the tutor.

# Greetings!

- Be welcome in this lecture hall!
- Please ask questions/let me know if I'm difficult to understand.
- This is an introduction to financial data analysis using Python.
  - The order matters!
  - Assumes knowledge of basic python.
  - Does more object oriented stuff.

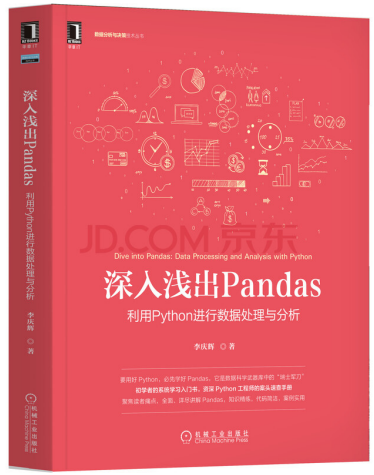
- Lecture & Laboratory Courses
  - 1-4,7-12,15 周二第 7-8 节，庄汉水楼（南强二）302
  - 5-6,13-14 周二第 7-8 节，保欣丽英楼（嘉庚一）512

# What will I be doing?

- Class participation (10%)
  - Attendance & Performance
- Assignments (30%)
  - 2 small assignments (2 * 5%)
    - Assignments must be done on your own.
    - Due 11:59 pm on due date, submitted in DingTalk.
    - The first assignment is meant to be small, it will be posted at week 3.
  - 2 projects (2 * 10%)
    - Projects will be posted at laboratory courses.
    - Projects can be done in pairs.
- Final exam (60%)
  - It will be closed book written test.

# The Book

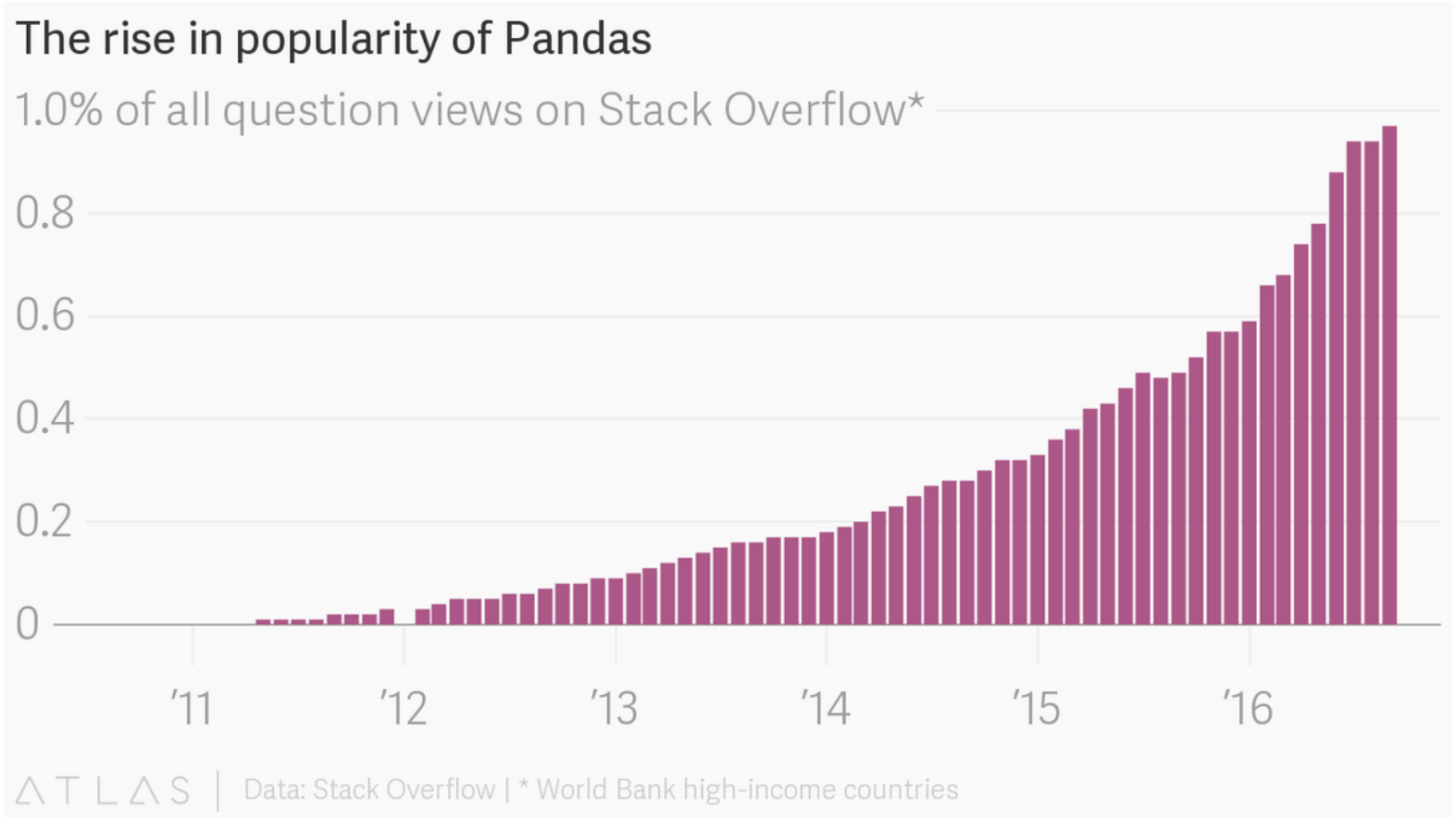
- 深入浅出 Pandas - 利用 Python 进行数据处理与分析 (李庆辉)




- Online resources
  - Pandas 教程 (李庆辉/盖若的官方网站)
  - Python 教程 (廖雪峰的官方网站)
  - Python for Data Analysis (by Wes McKinney)
  - Pandas official user guide (pandas documentation)
  - Stackoverflow, CSDN, GitHub

# Pandas questions on Stack Overflow

- Make full use of online resources.



# Academic Offences

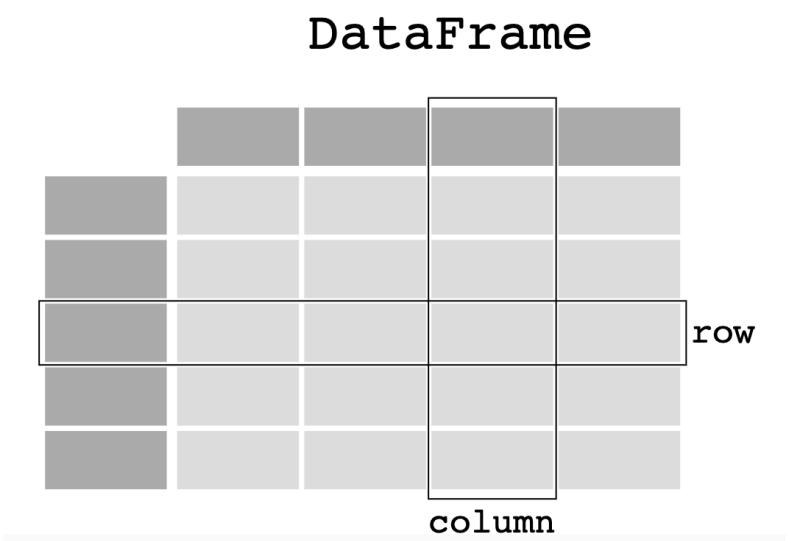
- You should do all the work that you submit.
  - Work by your project partner counts.
  - Never look at another teams works.
  - Never show another team your work.
  - Applies to all drafts and partial solutions.
  - Discuss how to solve an assignment only with course staff.

# Course Information

Q & A

# What kinds of data?

- The primary focus is on **structured data**.
  - Tabular or spreadsheet-like data.
    - Financial statements (e.g. balance sheet)
  - Multiple tables of data interrelated by key columns.
    - Key: firm name, stock code, ISIN
- Users of Microsoft Excel will not be strangers to these kinds of data.
  - A large percentage of datasets can be transformed into a structured form.



# Why Python for data analysis?

- Scripting language
- Features:
  - Easy-to-learn: relatively few keywords and simple structure.
  - Easy-to-read: clearly defined syntax and visible to the eyes.
  - Cross-platform compatible on Linux, Windows, and Macintosh.
  - Large and active scientific computing and data analysis **community**.
- Applications:
  - Data collection (urllib, request, selenium)
  - Data cleaning (**pandas**)
  - Data analysis (**pandas**, NumPy, matplotlib, scikit-learn, statsmodels)

# Introduction of Pandas

- What is Pandas?
  - Pandas is an **open-source** library used for working with data sets.
    - In particular, it offers data structures and operations for manipulating numerical tables and time series.
  - The name is derived from the term "panel data"
    - Observations over multiple time periods for the same individuals.
  - Developer: Wes McKinney
    - Researcher at *AQR Capital*, 2007-2010
    - For a flexible tool to perform quantitative analysis on financial data.

# What is Pandas for?

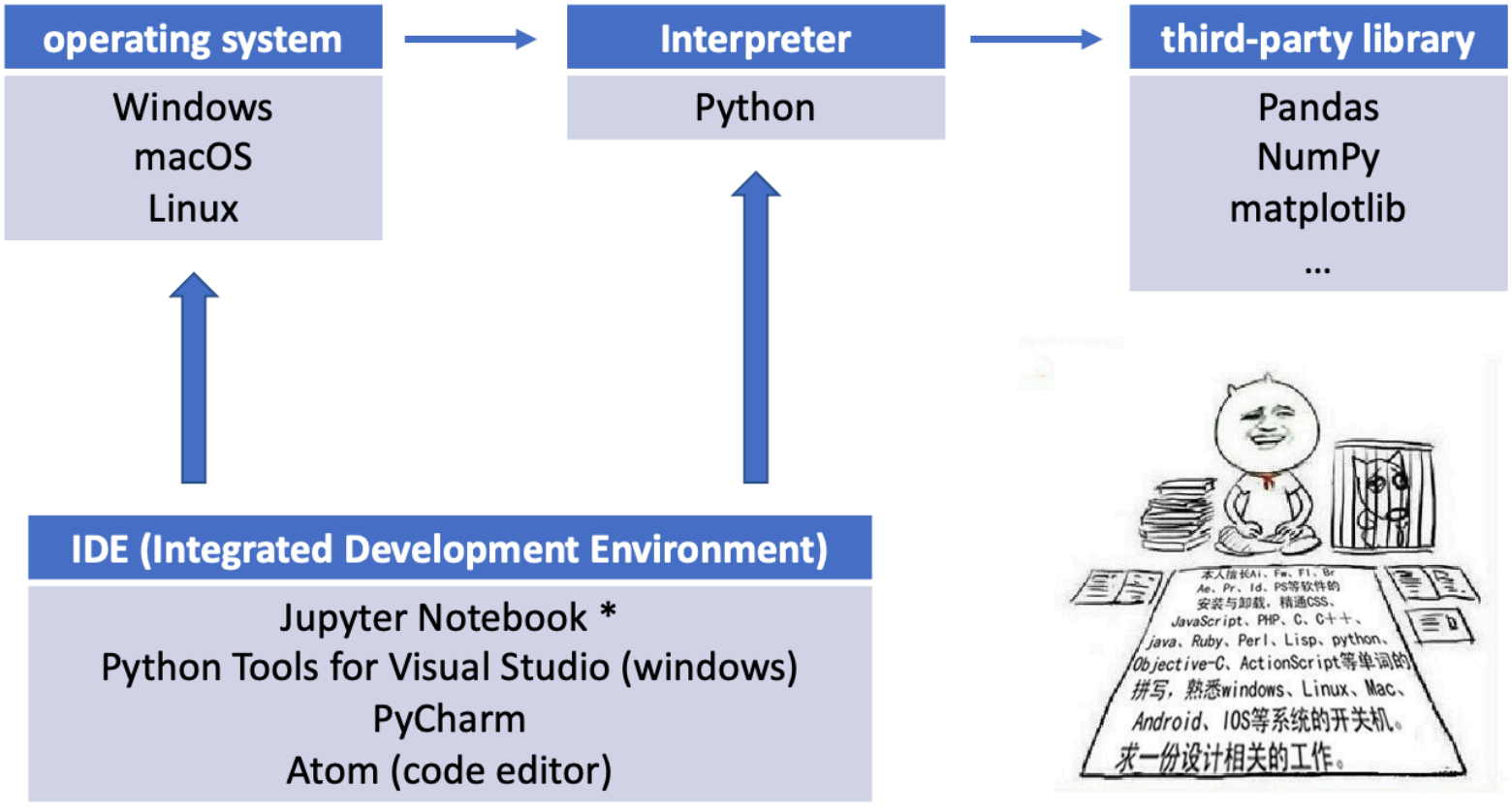
- 4 typical steps: load, clean, wrangling, and analyze.
  - Data loading and storage
    - Reading and writing data in multiple formats (**.csv** .xls .txt .json)
    - Indexing & reindexing
  - Data cleaning and preparation
    - Handling missing data
    - Data transformation
  - Data wrangling: join, combine, and reshape
    - Combining and merging datasets
    - Aggregation and group operations
  - Data analysis
    - Statistical analysis
    - Plotting and visualization

# What is Pandas for? 基本功能

- 从 Excel、CSV、网页、SQL、剪贴板等读取数据
- 合并多个文件或者 sheet 数据，拆分数据为独立文件
- 数据清洗，如去重、缺失值、填充默认值、格式补全、极端值处理等
- 建立高效的索引
- 支持大体量数据
- 按一定业务逻辑插入计算后的列、删除列
- 灵活方便的数据查询、筛选
- 分组聚合数据，可独立指定分组后的各字段计算方式
- 数据的转置，如行转列列转行变更处理
- 连接数据库，直接 SQL 查询数据并进行处理
- 对时序数据进行分组采样，如按月、按季、按工作小时，也可以自定义周期，如工作日
- 窗口计划，移动窗口统计、日期移动等
- 灵活的可视化图表输出，支持所有的统计图形
- 融合在表格的样式风格，提高数据识别效率

# Installation and Setup

# Python working environment



# miniconda

- miniconda
  - https://docs.conda.io/en/latest/miniconda (官方)
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/miniconda (国内镜像)

| | | |
|---|---|---|
| Miniconda3-py39_4.11.0-MacOSX-x86_64.pkg | 61.9 MiB | 2022-02-16 03:08 |
| Miniconda3-py39_4.11.0-Linux-x86_64.sh | 72.2 MiB | 2022-02-16 03:08 |
| Miniconda3-py39_4.11.0-MacOSX-x86_64.sh | 55.2 MiB | 2022-02-16 03:08 |
| Miniconda3-py39_4.11.0-Windows-x86.exe | 66.5 MiB | 2022-02-16 03:08 |
| Miniconda3-py39_4.11.0-Windows-x86_64.exe | 70.4 MiB | 2022-02-16 03:08 |

# Terminal

- Terminal (installation complete)
  - windows: 菜单或者桌面找到终端管理器 (Anaconda Prompt)
  - macOS: 启动器找到终端 (Termincal)

```
mac:
Last login: Tue Feb 22 00:46:09 on ttys000
(base) luping@Yus-MacBook-Pro ~ %

windows:
(base) PS C:\Users\luping>_
```

# Install third-party libraries

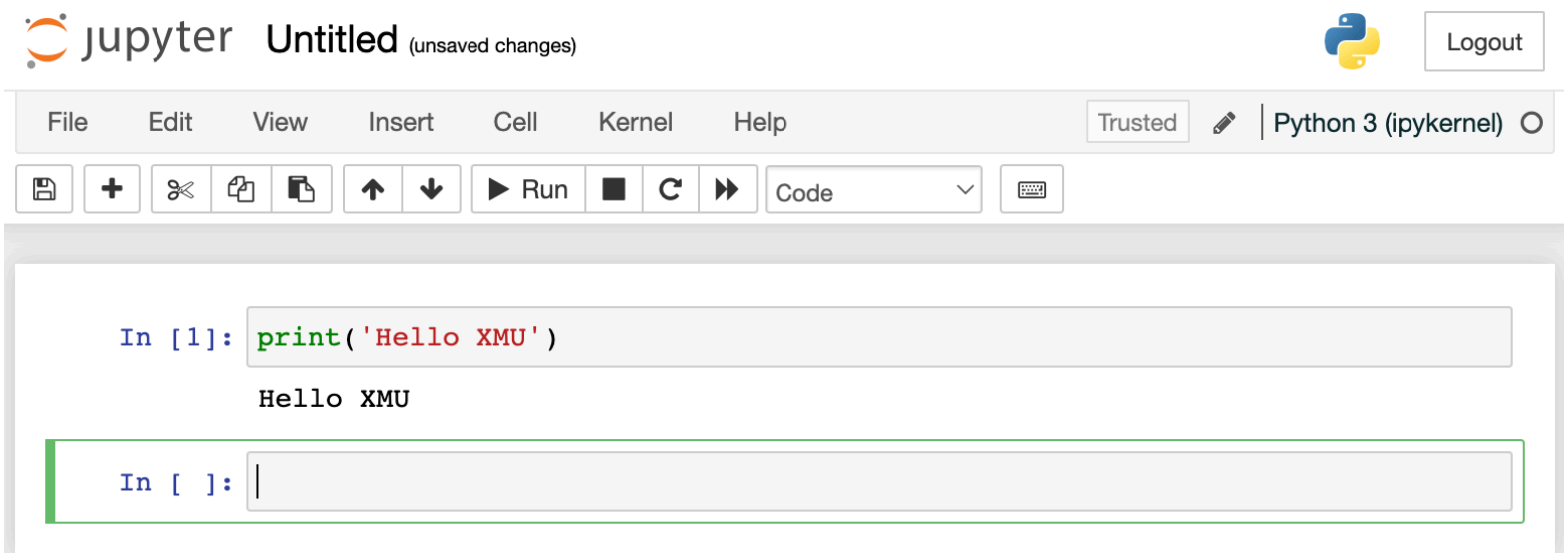
- pip: package installer
  - pip list: 查看当前 Python 环境安装了哪些库
  - pip install 库名: 安装新库
  - pip install 库名 -U: 升级库至最新版本
  - pip unintall 库名: 卸载库

# IDE: Jupyter Notebook

```
# 安装 Jupyter Notebook
pip install notebook

# 镜像安装 Jupyter Notebook (直接安装不成功时使用镜像安装)
pip install jupyter -i https://pypi.tuna.tsinghua.edu.cn/simple

#启动 Jupyter Notebook
jupyter notebook
```



# pandas

```
# 安装 pandas
pip install pandas

# 镜像安装 pandas（直接安装不成功时使用镜像安装）
pip install pandas -i https://pypi.tuna.tsinghua.edu.cn/simple

# 其他常用库（将上边代码中的 pandas 替换成以下包名进行安装）
# excel 处理相关包 xlrd openpyxl xlsxwriter
# 解析网页包 requests lxml html5lib BeautifulSoup4
# 可视化包 matplotlib seaborn plotly bokeh
# 计算包：scipy statsmodels
```