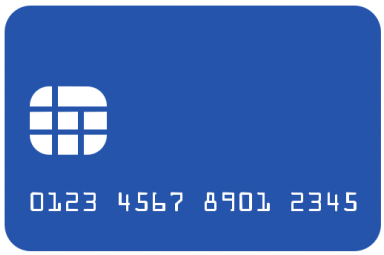
ETL

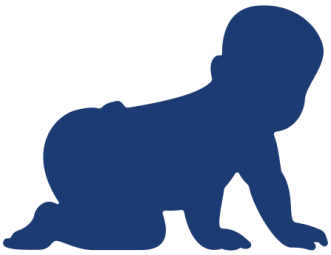
Extract:

第一個是資料的萃取,研究股市一般分為四個面向,基本面、技術面、籌碼面、消息面，前三者為結構化資料，資料來源除了馬上能聯想到的台灣證券交易所，也有許多民間經營的網站，而我們第一個碰到的問題，就是直接刮取證券交易所的資料會被網頁阻擋。

此時，我們解決的第一個辦法是找到了政府所提供的技術面與籌碼面的資料API，將Json格式的資料經由Python萃取後，再放入我們的Mysql。

另一方面，基本面的資料我們則找到了一個名為goodinfo的網站,這個網站所提供的資料原本是能直接使用網頁介面手動下載，因此我們先試著用了selenium來模仿人類點擊的動作，不久就馬上發現，這樣的效率實在太慢了。決定改用原本的方法，直接刮取網頁資料，此時一樣碰到了大量索要資料造成被網站阻擋，我們用了四種方法避開這個情況，一是使用在網路上找到的虛擬IP位址，利用不同的IP繞開網站的判定、再來是帶入真正使用網頁時的Headers與Cookies，使站方以為是真正的使用者再點擊網頁，最後則是在每個Request之間Sleep數秒，也是為了模仿真正人類使用的情況。





IP Headers Cookies

Transform：

解決的資料萃取的問題，再來是ETL的T，資料轉換的問題，無倫是基本面、籌碼面、技術面，同一公司的財報資料在不同年份可能就有不同的欄位名稱，不同公司之間更是有許多完全不一樣的資料欄位，上市與上櫃公司又是從不同的網頁所刮取，但是在最後，我們合併到MYSQL時會變成一個面向一張表格。

基本面的部分，我們主要是進行了缺值處理，再比對需要的欄位之後，特定公司出現的特定欄位予以刪除，而重要有缺損的公司則補上0來替代空值。

技術面與籌碼面的部分，我們則是先分別刮取台灣證券交易所和櫃買中心的上市與上櫃資料後，比對出他們之間不同名稱但是意義相同的欄位，再予以合併。

本業獲利	2015Q1	
	金額	%
銷貨收入淨額	-	-
營業收入	2,220	100
銷貨成本	-	-
營業成本	1,126	50.7
營業毛利	1,094	49.3
未實現銷貨損益	0.2	0.01
已實現銷貨損益	-	-
營業毛利淨額	1,094	49.3
推銷費用	13.91	0.63
管理費用	43.66	1.97
研究發展費用	167.8	7.56
營業費用	225.4	10.2

本業獲利	2020Q1	
	金額	%
營業收入	3,106	100
營業成本	1,498	48.2
營業毛利	1,608	51.8
未實現銷貨損益	-	-
已實現銷貨損益	-0.077	0
營業毛利淨額	1,608	51.8
推銷費用	14.51	0.47
管理費用	59.03	1.9
研究發展費用	249.7	8.04
營業費用	323.2	10.4
其他收益及費損合計	0.68	0.02
營業利益	1,285	41.4

Load：

資料萃取、轉換完成後，才到了大數據的問題核心:如何使用資料?
如何成數以萬計、甚至是億計得資料中找出它們之間的”關聯”，創造資料的價值，並加工成淺顯易懂才是大數據的精髓。

此時，我們的小組成員再使用其中股票每日的開、高、低、收、量來建模與畫圖分析時，才發現原本使用的資料來源近年雖然正確，但是時間放到2000年左右則有很誇張的錯誤(例:一天的交易量(張數)破百萬甚至千萬)，在上課期間，從每個老師口中最常聽到:Garbage In，Garbage Out。使用錯誤的資料來源，也只會得到完全沒有價值的資訊，因此我們才緊急再從不同的地方，重新刮取了正確的資料，才得以好好利用它們。

而使用的部分呢，我分配到的是如何將股票資料變身成為投資人最愛看的圖片，無倫是量價背離、黃金交叉、多頭，許許多多的股民判斷如何投資時，常常只需要幾張圖片便能看出大量資訊，因此我們使用Python從我們剛剛刮取得資料來畫成許多常用圖片，如:K線圖、KD線圖、MACD圖 等，並結合後面會說到的Linebot部分，在使用者提出需求時，來將他所需要的內容回傳，才實現真正”使用”資料。

主講人：蘇于傑

https://github.com/jaysu99-lab