NBA data:
http://stat-nba.com/navigator.php
從(00-01)球季開始有上下半場的紀錄，18年、一年82場比賽，共1476場比賽。
目前想依據上半場的籃板數、投籃命中率、三分命中率、罰球數、犯規、失誤、助攻以及得分做預測，共八種featurs.

盤口數據:
https://www.playsport.cc/predictgame.php?action=result&allianceid=3&gametime=20131101

社群網路分析
-趨勢分析
-議題影響力
-用戶關係預測
>分析結果用途？

Keras word2vector的教學:
http://gavinhuang.github.io/blog/keras-based-word2vec-tutorial/

報告
https://drive.google.com/open?id=1_P-geXMqBHBTkSR6S83TNeDlokTJ7RJO

Final project:

nlp相關例子:
https://towardsdatascience.com/using-nlp-and-deep-learning-to-predict-the-stock-market-64eb9229e102

有關股票預測模型選擇的PAPER:
https://drive.google.com/drive/folders/1KBSFRRsg_on5-7fOEQe4FlGejzZuOmAk

Sentece2vector:
https://tfhub.dev/google/universal-sentence-encoder/2
https://medium.com/huggingface/universal-word-sentence-embeddings-ce48ddc8fc3a

Glove:
https://blog.csdn.net/sinat_26917383/article/details/54847240
https://blog.csdn.net/u014665013/article/details/79642083
https://ithelp.ithome.com.tw/articles/10194633

Word to Sentence meaning:
https://nlp.stanford.edu/manning/talks/Simons-Institute-Manning-2017.pdf

example of feature importance :

https://www.kaggle.com/bguberfain/a-simple-model-using-the-market-and-news-data/notebook

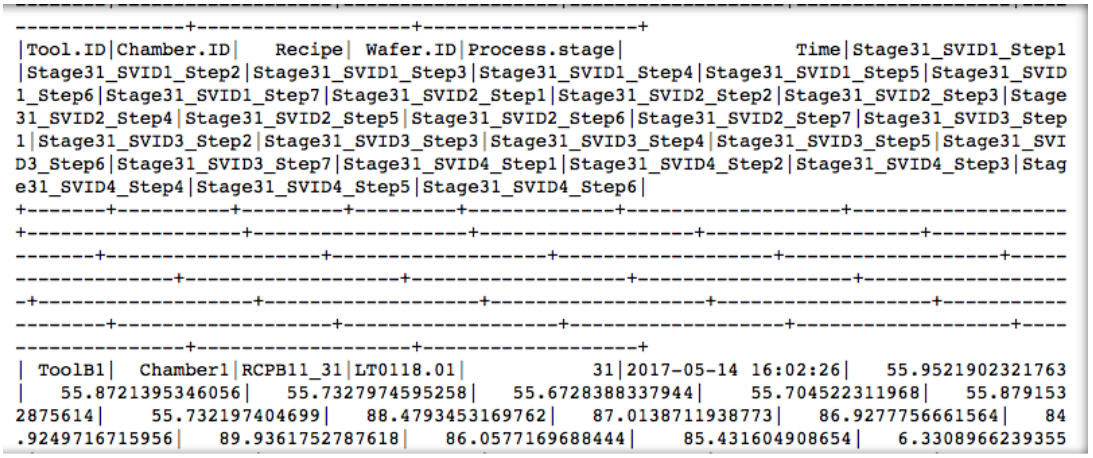
---

期中報告
資料類別:

wafer_information/Wafer_log
　　有WAT.ID Process.stage Process Tool.ID Time Action
　　主要紀錄wafer 在哪一個階段的哪一個Tool, 何時進入LOGIN, 何時LOGOUT
Yield裡的
LT0001.01 代表的意思? (第一批的第一片？)

	yield
LT0001.01	76.97957
LT0001.02	75.20292
LT0001.03	77.02167
LT0001.04	75.20238

HH/FDC_data/FDC_Stage1.csv



```
--------------+------------------+------------------+
|Tool.ID|Chamber.ID|    Recipe| Wafer.ID|Process.stage|              Time|Stage31_SVID1_Step1
|Stage31_SVID1_Step2|Stage31_SVID1_Step3|Stage31_SVID1_Step4|Stage31_SVID1_Step5|Stage31_SVID
1_Step6|Stage31_SVID1_Step7|Stage31_SVID2_Step1|Stage31_SVID2_Step2|Stage31_SVID2_Step3|Stage
31_SVID2_Step4|Stage31_SVID2_Step5|Stage31_SVID2_Step6|Stage31_SVID2_Step7|Stage31_SVID3_Step
1|Stage31_SVID3_Step2|Stage31_SVID3_Step3|Stage31_SVID3_Step4|Stage31_SVID3_Step5|Stage31_SVI
D3_Step6|Stage31_SVID3_Step7|Stage31_SVID4_Step1|Stage31_SVID4_Step2|Stage31_SVID4_Step3|Stag
e31_SVID4_Step4|Stage31_SVID4_Step5|Stage31_SVID4_Step6|
+-------+----------+---------+---------+-------------+------------------+------------------
+------------------+------------------+------------------+------------------+------------
-------+------------------+------------------+------------------+------------------+-----
-------------+------------------+------------------+------------------+------------------
-+------------------+------------------+------------------+------------------+------------
--------+------------------+------------------+------------------+------------------+----
--------------+------------------+------------------+
| ToolB1|  Chamber1|RCPB11_31|LT0118.01|           31|2017-05-14 16:02:26|   55.9521902321763
|   55.8721395346056|   55.7327974595258|   55.6728388337944|    55.704522311968|   55.879153
2875614|    55.732197404699|   88.4793453169762|   87.0138711938773|   86.9277756661564|   84
.9249716715956|   89.9361752787618|   86.0577169688444|    85.431604908654|   6.3308966239355
```

集束型製程設備的狀態變數編號(SVID):
　　連續不間斷的製程變數稱之為狀態變數, 例如濕度、風速、排氣、熱板溫度、時間等等。

是否能找出在每個step製成變數對於每個svid的median gap 作為一個feature?

SVID1和SVID2差別?

ANS:機台類型
SVID1_Step1和SVID1_Step2差別?

HW1:
WAT vs Yield (pearson correlation)

HW2:
wafer_information: process stage --tool id (sort by median gap with yield)

idea pitch:
每個stage中不同svid的平均數與yield的相關性

All of the parameters we have so far:

wat_data/ parameter_set

```
waf_para.show()
```

```
+---+------+---------+
|_c0|    aa|    Range|
+---+------+---------+
|  1|0.3641| 150000.0|
|  2|0.2456|  30000.0|
|  3|0.4573|    500.0|
|  4|0.4453|2000000.0|
|  5|0.4706|      1.0|
|  6|0.1531|      1.0|
|  7|0.2258|     0.05|
|  8|0.3396|     20.0|
|  9| 0.177|    500.0|
| 10|0.1178|     20.0|
| 11|0.4162|  30000.0|
| 12|0.4788|      1.0|
| 13|0.1953|   1.0E-4|
| 14|0.3469|   1.0E-4|
| 15|0.3046|  30000.0|
| 16|0.2257|    500.0|
| 17|0.4291| 150000.0|
| 18|0.1553|      1.0|
| 19|0.4791|      1.0|
| 20| 0.355|   1000.0|
+---+------+---------+
only showing top 20 rows
```

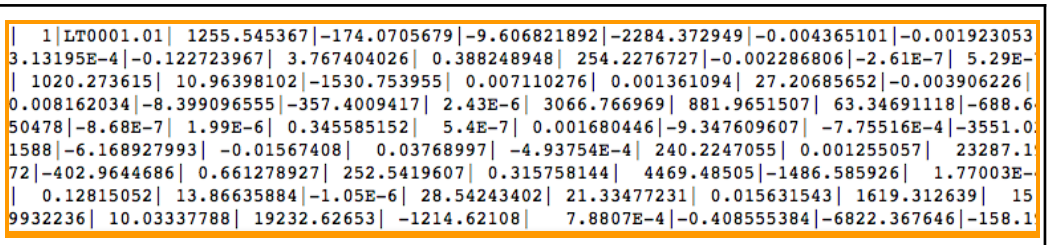
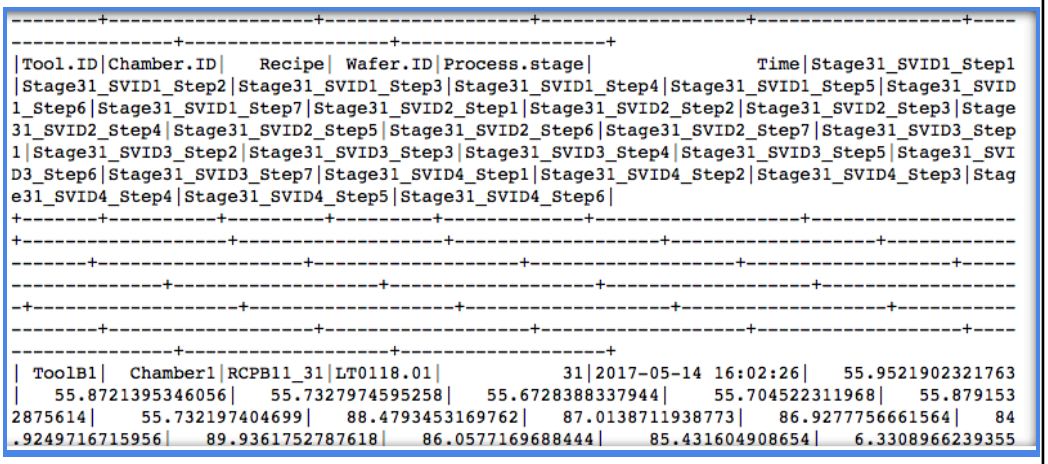
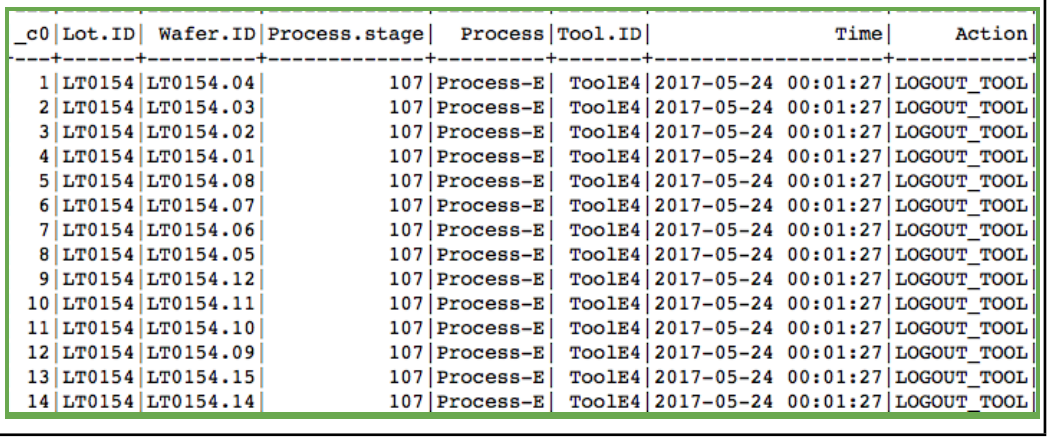
wat_data/ wat_root_cause

```
waf_root.show()
```

```
+---+----------+
|_c0|root_cause|
+---+----------+
|  1|      2064|
|  2|      1477|
|  3|      1036|
|  4|      2985|
|  5|      2086|
|  6|        33|
|  7|      1036|
|  8|       517|
|  9|      2848|
| 10|       748|
+---+----------+
```

wat_data/ wat

```
|_c0| wafer_id|      WAT1|      WAT2|      WAT3|      WAT4|      WAT5|      WAT6|
WAT7|      WAT8|      WAT9|     WAT10|     WAT11|     WAT12|  WAT13|  WAT14|
WAT15|     WAT16|     WAT17|     WAT18|     WAT19|     WAT20|     WAT21|     WA
T22|     WAT23|     WAT24|  WAT25|     WAT26|     WAT27|     WAT28|     WAT29|
WAT30|  WAT31|     WAT32|  WAT33|     WAT34|     WAT35|     WAT36|     WAT37|
WAT38|     WAT39|     WAT40|     WAT41|     WAT42|     WAT43|     WAT44|     WA
T45|     WAT46|     WAT47|     WAT48|     WAT49|     WAT50|     WAT51|     WAT5
2|     WAT53|  WAT54|     WAT55|     WAT56|     WAT57|     WAT58|     WAT59|
WAT60|     WAT61|     WAT62|     WAT63|     WAT64|     WAT65|     WAT66|     WA
T67|     WAT68|     WAT69|  WAT70|     WAT71|     WAT72|     WAT73|     WAT74|
```

	
FDC_data/stage ~~	
wafer_informatio n/ wafer log	

Yield	+---------+----------------+ | _c0| yield| +---------+----------------+ |LT0001.01|76.9795724366398| |LT0001.02|75.2029172318689| |LT0001.03|77.0216726433812| |LT0001.04|75.2023758930297| |LT0001.05|79.2773721175721| |LT0001.06|77.2425479387875| |LT0001.07|76.4054086079774| |LT0001.08| 78.217875571929| |LT0001.09|75.5875504902877| |LT0001.10|75.2420699261284| |LT0001.11|75.3914766769037| |LT0001.12|76.1083246485199| |LT0001.13|72.2381387653298| |LT0001.14|75.9324870525181| |LT0001.15|80.6229895244045| |LT0001.16| 77.944680338383| |LT0001.17|77.9292160824658|

yield	wat	_c0 = wafer_id
yield	wafer_log	_c0 = wafer_id
yield	fdc_data	_c0 = wafer_id

features:

fab:
average/median of svid in every process stage

wat:
root_cause, parameter_set, watt

檔名和裡面的stage不一樣的有
stage1,13 ,18都是31

---

http://140.116.158.43:8889/notebooks/drop%20NA.ipynb程式碼
FDC_Stage1=FDC_Stage1.where(trim(FDC_Stage1.Stage31_SVID1_Step1)!='NA')
我把每一個stage的csv檔中選擇SVID_Step1 不為NA的值挑出來
這樣應該就是會去除chamber NA的值

FDC_Stage10.write.csv('HH/stage/FDC_Stage10.csv')
這個是存在hdfs裡面的stage資料夾裡都東西
如果需要讀的話應該可以從這邊讀乾淨的資料

合併檔在hdfs裡HH/stage/mid2_80.csv　mid81_200.csv　mid201_300.csv　共三個
join的話要用 'waferID'

```
from pyspark.ml import Pipeline
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.feature import IndexToString, StringIndexer, VectorIndexer
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

# Load and parse the data file, converting it to a DataFrame.
data = spark.read.format("libsvm").load("data/mllib/sample_libsvm_data.txt")

# Index labels, adding metadata to the label column.
# Fit on whole dataset to include all labels in index.
labelIndexer = StringIndexer(inputCol="label", outputCol="indexedLabel").fit(data)

# Automatically identify categorical features, and index them.
# Set maxCategories so features with > 4 distinct values are treated as continuous.
featureIndexer =\
    VectorIndexer(inputCol="features", outputCol="indexedFeatures", maxCategories=4).fit(data)

# Split the data into training and test sets (30% held out for testing)
(trainingData, testData) = data.randomSplit([0.7, 0.3])

# Train a RandomForest model.
rf = RandomForestClassifier(labelCol="indexedLabel", featuresCol="indexedFeatures", numTrees=10)

# Convert indexed labels back to original labels.
labelConverter = IndexToString(inputCol="prediction", outputCol="predictedLabel",
                               labels=labelIndexer.labels)

# Chain indexers and forest in a Pipeline
pipeline = Pipeline(stages=[labelIndexer, featureIndexer, rf, labelConverter])

# Train model.  This also runs the indexers.
model = pipeline.fit(trainingData)
```

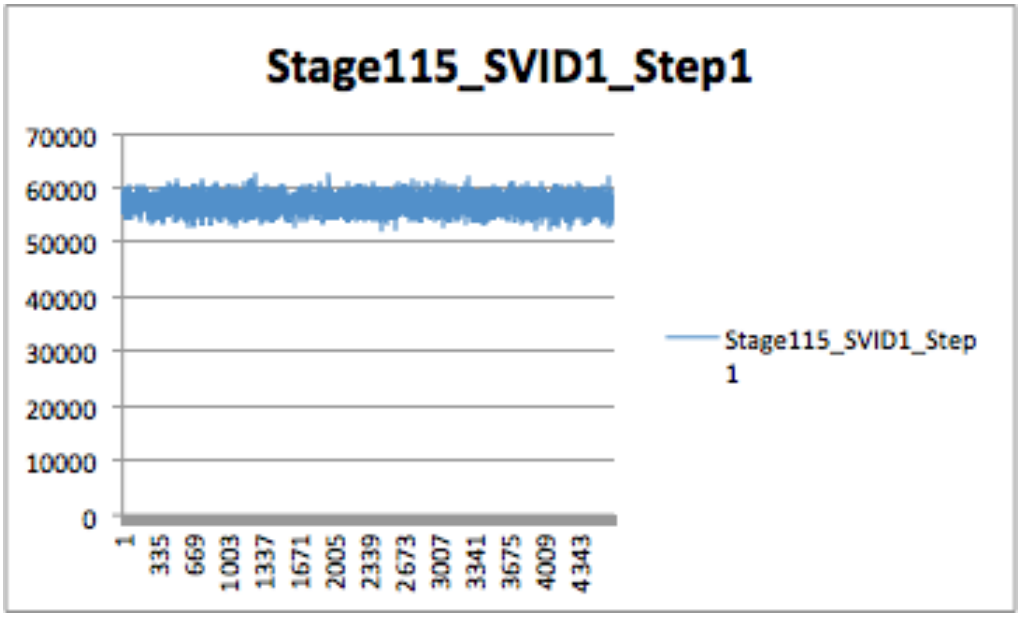
WAT using PLSR

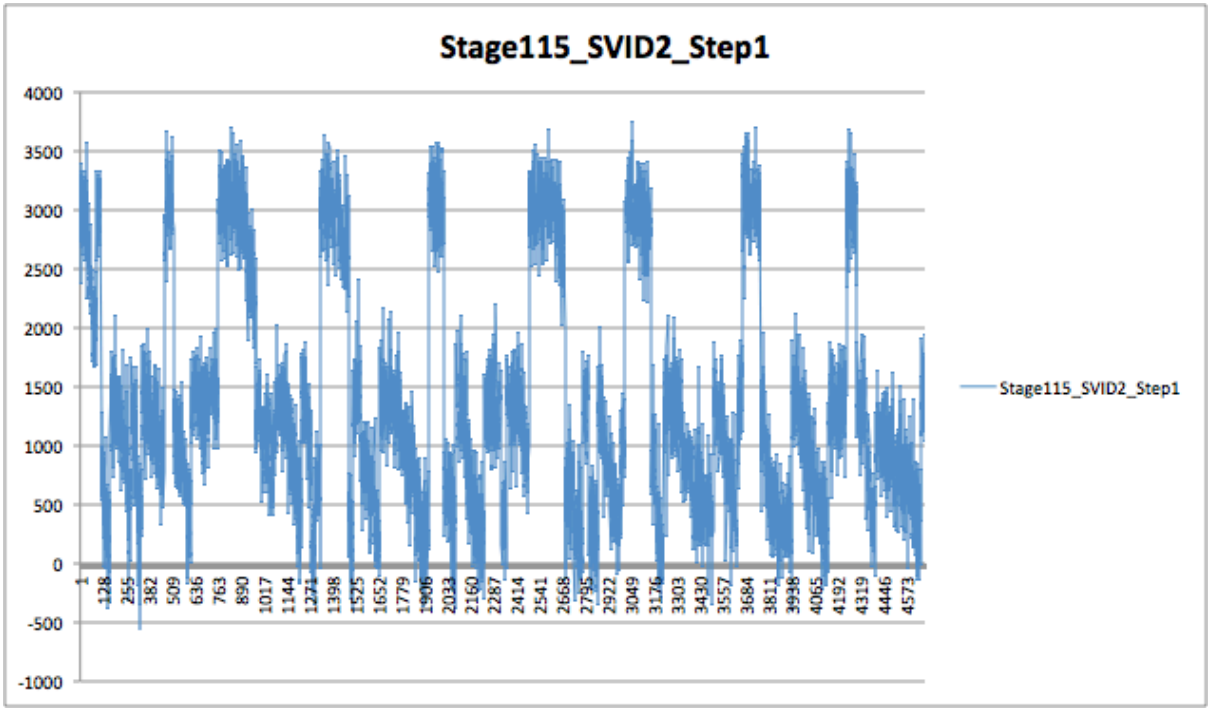
TOP 20 VIP score:

('WAT1036', 2.483739861057725)
('WAT2985', 2.422242049574934)
('WAT2848', 2.1990348543163294)
('WAT748', 2.0482288785042795)
('WAT517', 1.976433769145779)
('WAT1477', 1.9581630955602245)
('WAT33', 1.952304726606573)
('WAT2064', 1.9073156190404252)
('WAT2086', 1.881534819170371)
('WAT1707', 1.549297648424166)
('WAT2871', 1.5491255422894321)
('WAT1920', 1.5489650886320314)
('WAT845', 1.5483752552917456)
('WAT2768', 1.5480766672322124)
('WAT882', 1.5477374304674367)
('WAT710', 1.5470169628989725)
('WAT308', 1.546955073607509)
('WAT116', 1.5464957942353468)
('WAT2336', 1.5462959677894859)
('WAT2450', 1.5451915632394932)

yield	1.000000
Stage7_SVID2_Step3	0.123576
Stage7_SVID2_Step6	0.123264
Stage7_SVID2_Step9	0.123057
Stage7_SVID2_Step7	0.122793
Stage7_SVID2_Step5	0.122508
Stage7_SVID2_Step2	0.122426
Stage7_SVID2_Step1	0.122335
Stage7_SVID2_Step4	0.122320
Stage7_SVID2_Step8	0.121990
Stage5_SVID3_Step4	0.082902
Stage5_SVID2_Step1	0.076705
Stage5_SVID3_Step5	0.075878
Stage5_SVID3_Step6	0.075223
Stage5_SVID3_Step3	0.073265

Stage115_SVID2_Step5	2.259955
Stage115_SVID2_Step2	2.249139
Stage115_SVID2_Step6	2.248554
Stage115_SVID2_Step3	2.244098
Stage115_SVID2_Step1	2.243292
Stage115_SVID2_Step4	2.209557
Stage278_SVID6_Step2	1.848276
Stage278_SVID6_Step5	1.838912
Stage278_SVID6_Step1	1.832219
Stage278_SVID6_Step3	1.828579
Stage278_SVID6_Step4	1.809828
Stage228_SVID2_Step3	1.739759
Stage228_SVID2_Step5	1.714077
Stage228_SVID2_Step8	1.713799
Stage228_SVID2_Step4	1.711585
Stage228_SVID2_Step7	1.694825
Stage228_SVID2_Step6	1.694662
Stage228_SVID2_Step2	1.689721
Stage228_SVID2_Step1	1.681943
Stage162_SVID2_Step6	1.582366
Stage162_SVID2_Step3	1.581654
Stage162_SVID2_Step7	1.573495
Stage162_SVID2_Step9	1.569181
Stage162_SVID2_Step1	1.556815
Stage162_SVID2_Step4	1.551212
Stage289_SVID3_Step6	1.548174
Stage162_SVID2_Step2	1.543803
Stage289_SVID3_Step8	1.542024
Stage162_SVID2_Step8	1.538619
Stage289_SVID3_Step4	1.536764




Stage115_SVID2_Step1
4000
3500
3000
2500
2000
1500
1000
500
0
-500
-1000
1
128
255
382
509
636
763
890
1017
1144
1271
1398
1525
1652
1779
1906
2033
2160
2287
2414
2541
2668
2795
2922
3049
3176
3303
3430
3557
3684
3811
3938
4065
4192
4319
4446
4573
Stage115_SVID2_Step1