



# S3. Data Mining

# Caio Fernandes Moreno

# Problem Description

- A complex modern **semi-conductor manufacturing process** is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points.
- However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain **a combination of useful information, irrelevant information as well as noise.**

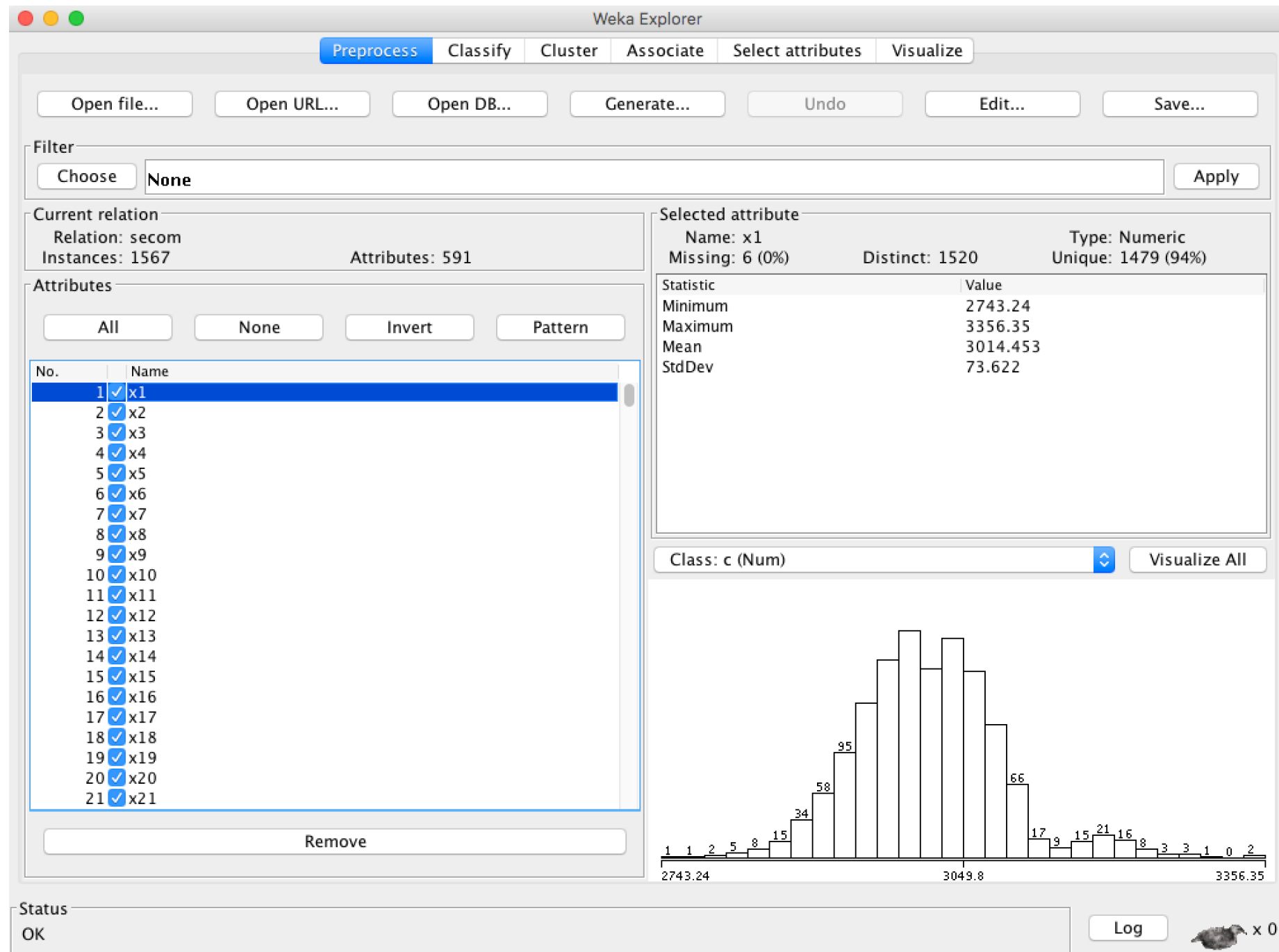
# Problem Description

- Engineers typically have a much larger number of signals than are actually required. **If we consider each type of signal as a feature**, then feature selection may be applied to identify the **most relevant signals**.
- The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs.

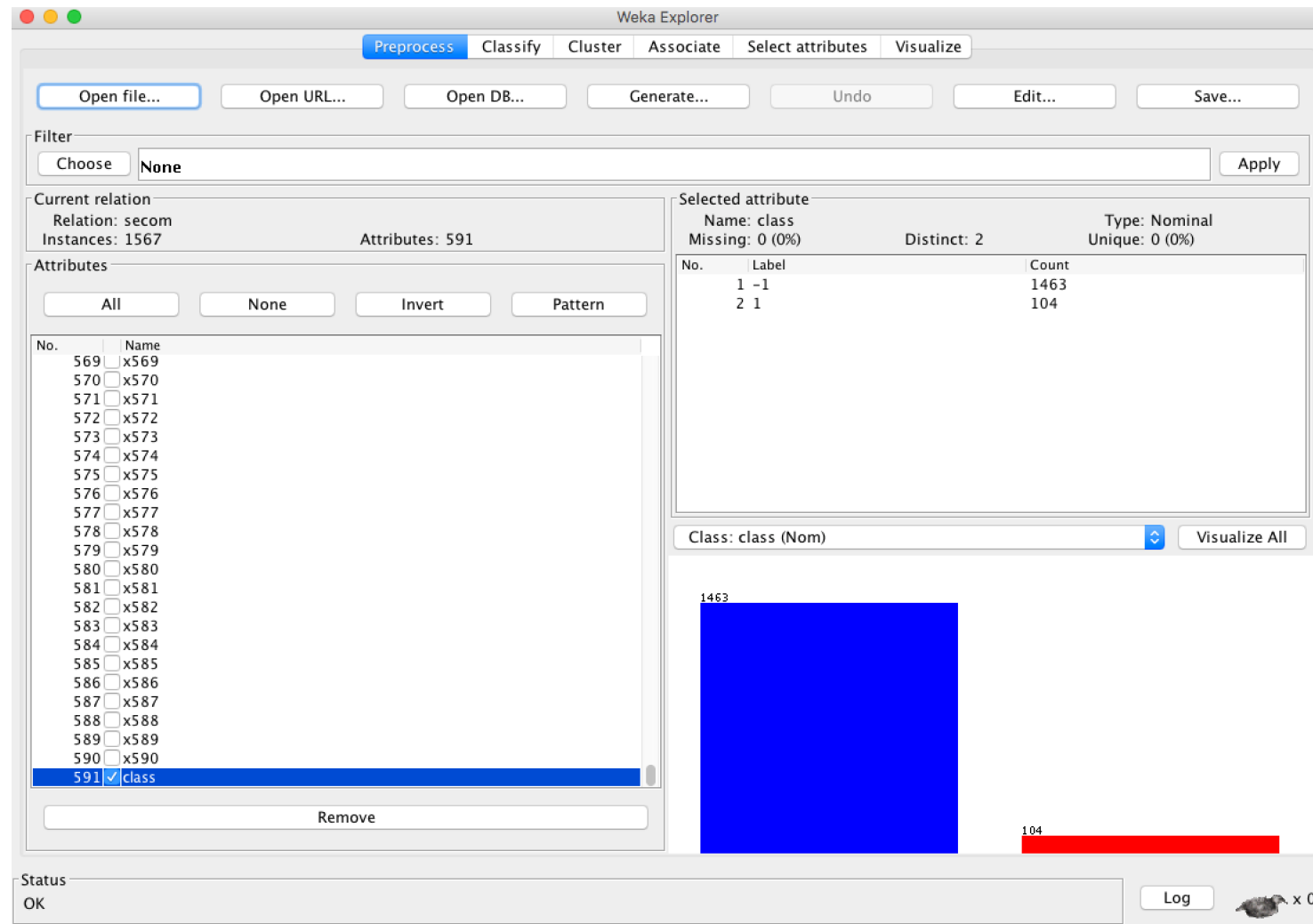
# SECOM Dataset

- SECOM Dataset: 1567 examples 591 features, 104 fails
- There are missing values;
- Where  $-1$  corresponds to a pass and  $1$  corresponds to a fail and the data time stamp is for that specific test point.
- <https://archive.ics.uci.edu/ml/datasets/SECOM>

# Features (variables)



# ARFF File



The variable **class** is the binary.

@attribute class {-1,1}

-1 = Pass  
1 = Fail

```
@attribute x588 numeric
@attribute x589 numeric
@attribute x590 numeric
@attribute class {-1,1}
```

```
@data
3030.93,2564,2187.7333,1411.1265,1.3602,100,97.61
455,202.4396,0,7.9558,414871,10.0433,968,192.3963
751,0,0055,1773,0010,04,0000,0,0000,0,1000,0,510
```

# Algorithms used

- Naive Bayes
- TAN
- IB1
- Idk
- RIPPER
- ID3
- C4.5 (J48)
- Logistic

# Measures

=== Stratified cross-validation ===  
 === Summary ===

Correctly Classified Instances	1012	64.582 %
Incorrectly Classified Instances	555	35.418 %
Kappa statistic	0.0077	
Mean absolute error	0.3534	
Root mean squared error	0.59	
Relative absolute error	283.9367 %	
Root relative squared error	237.0124 %	
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.666	0.644	0.936	0.666	0.778	0.505	-1
	0.356	0.334	0.07	0.356	0.118	0.497	1
Weighted Avg.	0.646	0.624	0.878	0.646	0.735	0.505	

=== Confusion Matrix ===

a	b	<-- classified as
975	488	a = -1
67	37	b = 1

		prediction outcome		total
		$p$	$n$	
actual value	$p'$	True Positive	False Negative	$P'$
	$n'$	False Positive	True Negative	$N'$
total		$P$	$N$	

It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative and that 64.58% of correctly classified instances.



# Naive Bayes with all variables.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      1012           64.582 %
Incorrectly Classified Instances    555           35.418 %
Kappa statistic                    0.0077
Mean absolute error                 0.3534
Root mean squared error             0.59
Relative absolute error             283.9367 %
Root relative squared error         237.0124 %
Total Number of Instances          1567

=== Detailed Accuracy By Class ===

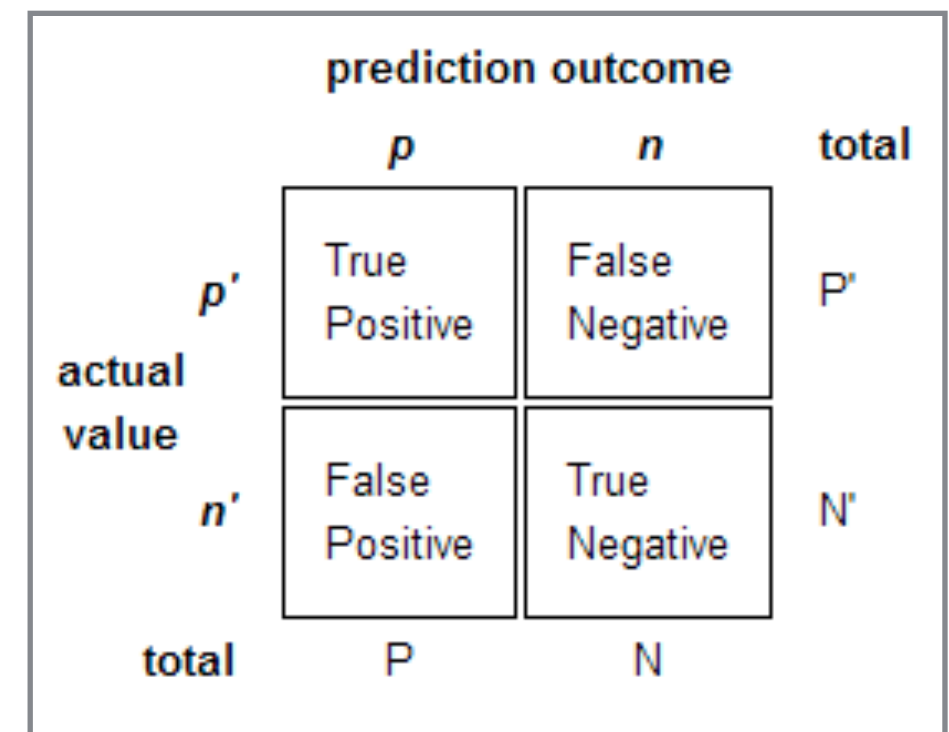
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.666	0.644	0.936	0.666	0.778	0.505	-1
	0.356	0.334	0.07	0.356	0.118	0.497	1
Weighted Avg.	0.646	0.624	0.878	0.646	0.735	0.505	

```

=== Confusion Matrix ===
  a  b  <-- classified as
975 488 |  a = -1
 67  37 |  b = 1

```



It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative. 64.58% Correctly Classified Instances.

# Try all algorithms

Using Weka I will try different algorithms using different techniques and compare the results in a table.

	All variables	FSS1	FSS2	Wrapper
Naive Bayes				
TAN				
IB1				
IBK				
RIPPER				
ID3				
C4.5 (J48)				
Logistic				

# BayesNet

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected. The classifier chosen is 'BayesNet' with the command: `-D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5`.

**Test options:**

- ☐ Use training set
- ☐ Supplied test set (Set...)
- ☒ Cross-validation (Folds: 10)
- ☐ Percentage split (%: 66)
- More options...

**Result list (right-click for options):**

- 00:54:48 - bayes.BayesNet

**Classifier output:**

```
x584(1): class
x585(1): class
x586(1): class
x587(1): class
x588(1): class
x589(1): class
x590(1): class
class(2):
LogScore Bayes: -8417.028179615854
LogScore BDeu: -8437.642984910786
LogScore MDL: -8456.471985605358
LogScore ENTROPY: -8364.510507575906
LogScore AIC: -8389.510507575906
```

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===  
=== Summary ===

Correctly Classified Instances	1402	89.4703 %
Incorrectly Classified Instances	165	10.5297 %
Kappa statistic	0.1689	
Mean absolute error	0.1309	
Root mean squared error	0.2765	
Relative absolute error	105.1604 %	
Root relative squared error	111.0764 %	
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.942	0.769	0.945	0.942	0.944	0.733	-1
	0.231	0.058	0.22	0.231	0.225	0.733	1
Weighted Avg.	0.895	0.722	0.897	0.895	0.896	0.733	

=== Confusion Matrix ===

a	b	<-- classified as	
1378	85	a = -1	
80	24	b = 1	

Status: OK

Log x 0

# Clusters (EM)



Clustered Instances

0	620 ( 40%)
1	532 ( 34%)
2	73 ( 5%)
3	161 ( 10%)
4	181 ( 12%)

Number of clusters: 5

Attribute	Cluster 0 (0.25)	1 (0.36)	2 (0.11)	3 (0.1)	4 (0.18)
x1					
mean	3008.3732	3017.0769	3017.2189	3007.9674	3019.6939
std. dev.	65.0659	78.9663	69.7525	78.5341	71.4975
x2					
mean	2495.3421	2495.192	2498.6908	2502.606	2492.1228
std. dev.	90.1282	71.9495	78.5174	83.8471	79.7755
x3					
mean	2202.0203	2201.1245	2201.2211	2198.8636	2197.8491
std. dev.	28.7128	29.2509	28.7702	32.5015	28.7759
x4					
mean	161408.8011	137357.8632	194804.6429	175739.5712	222154.6529
std. dev.	442682.2112	422197.3147	520856.5946	492155.3222	524988.1034
x5					
mean	146.0039	120.5253	128.7204	115.2781	97.4385
std. dev.	418.5318	404.4045	425.4729	417.1137	365.567
x6					
mean	100	100	100	100	100
std. dev.	0	0	0	0	0
x7					
mean	102.2575	100.0769	100.0893	101.651	101.9432
std. dev.	3.3735	7.3454	6.0552	5.8141	6.6374

weka.clusterers.EM

About

Simple EM (expectation maximisation) class.

More

Capabilities

debug ☐ False

displayModelInOldFormat ☐ False

maxIterations

minStdDev

numClusters

seed

Open... Save... OK Cancel

# Performance

	All variables	FSS1	FSS2	Wrapper
Naive Bayes	64.58%	75.62%	90.04%	<b>93.36%</b>
TAN	<b>93.23%</b>	<b>93.23%</b>	<b>93.29%</b>	<b>93.36%</b>
IB1	82.76%	72.68%	88.70%	84.74%
IBK	82.76%	72.68%	88.70%	<b>93.36%</b>
RIPPER	93.10%	<b>93.23%</b>	<b>93.29%</b>	<b>93.36%</b>
ID3	61.07%	65.15%	70.83%	<b>93.36%</b>
C4.5 (J48)	90.42%	92.40%	92.53%	<b>93.36%</b>
Logistic	80.91%	93.17%	<b>93.29%</b>	<b>93.36%</b>

# Thanks

- Caio Fernandes Moreno  
[caio.fmoreno@alumnos.upm.es](mailto:caio.fmoreno@alumnos.upm.es)