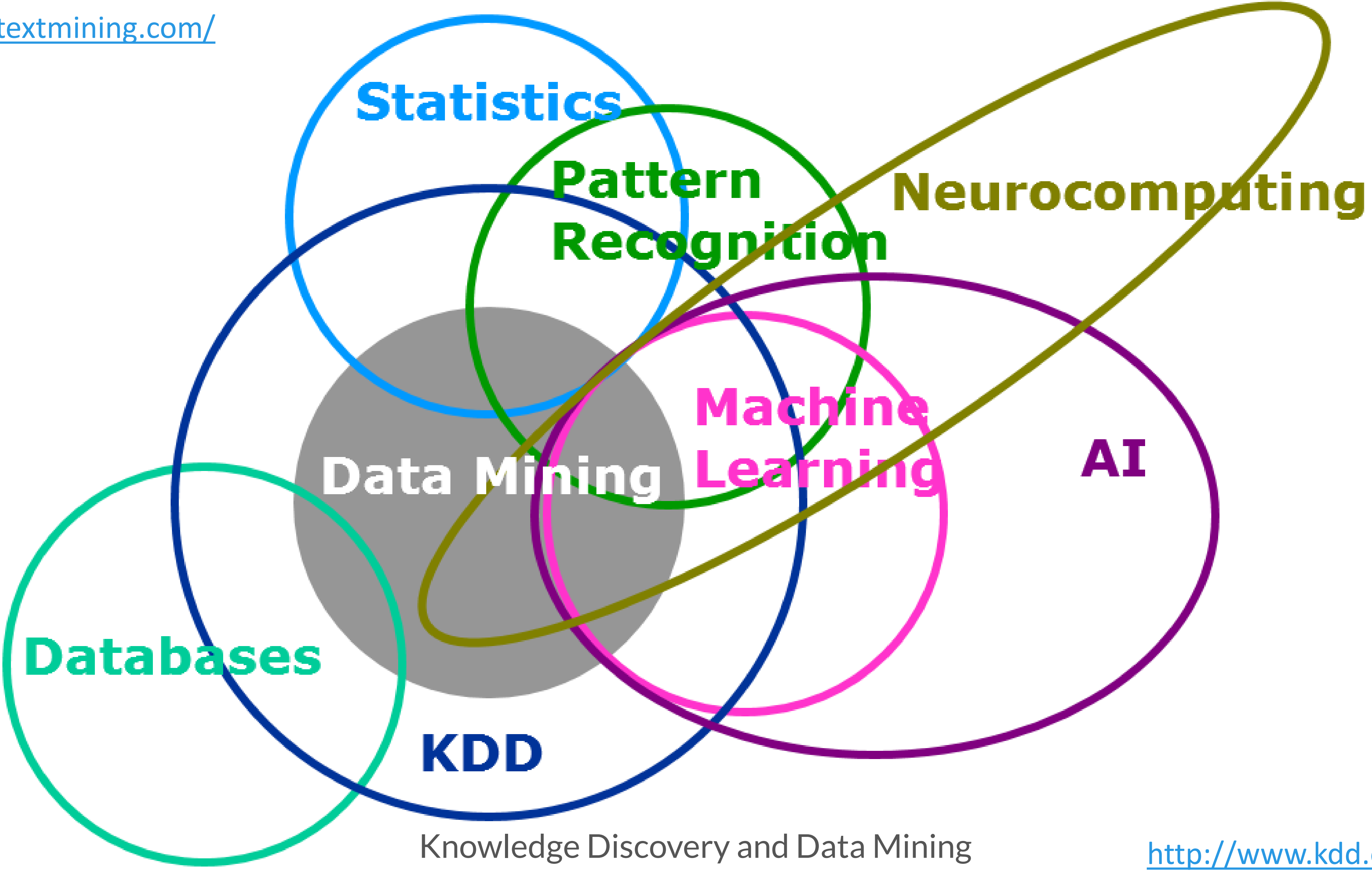
國立臺灣大學
National Taiwan University

資料科學與機器學習

國立臺灣大學共同教育中心

蔡芸琤

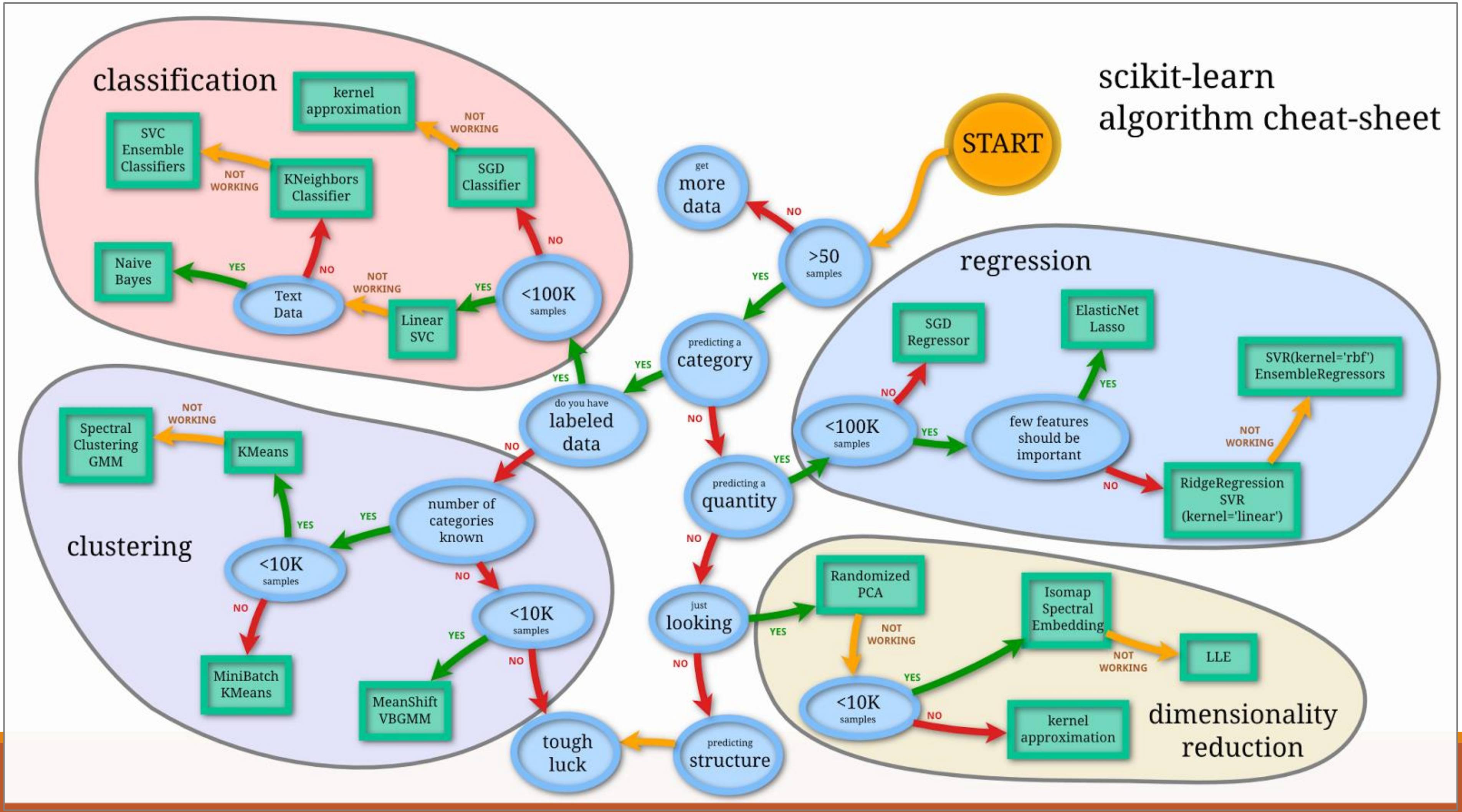
http://tidytextmining.com/

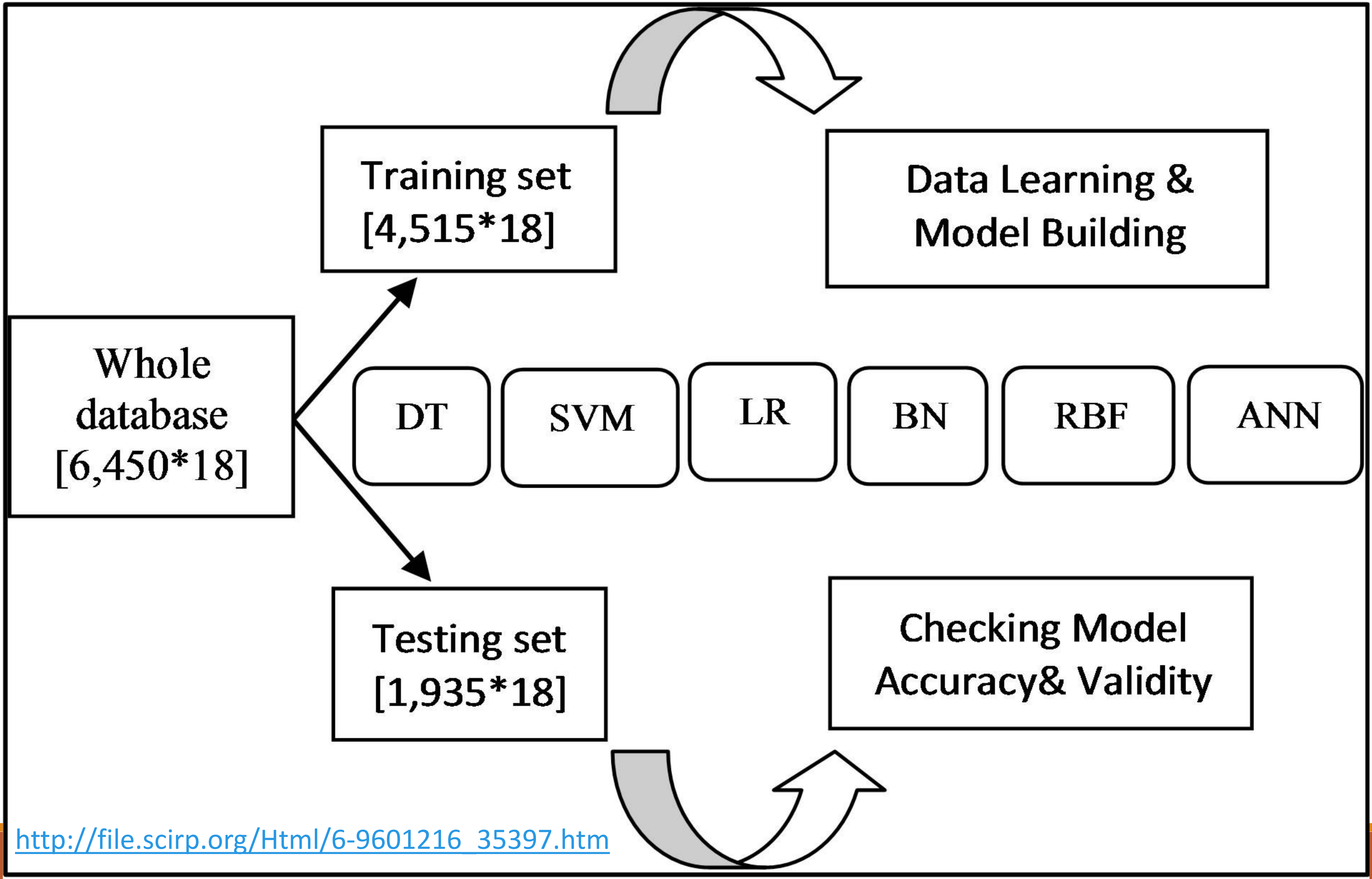


Knowledge Discovery and Data Mining

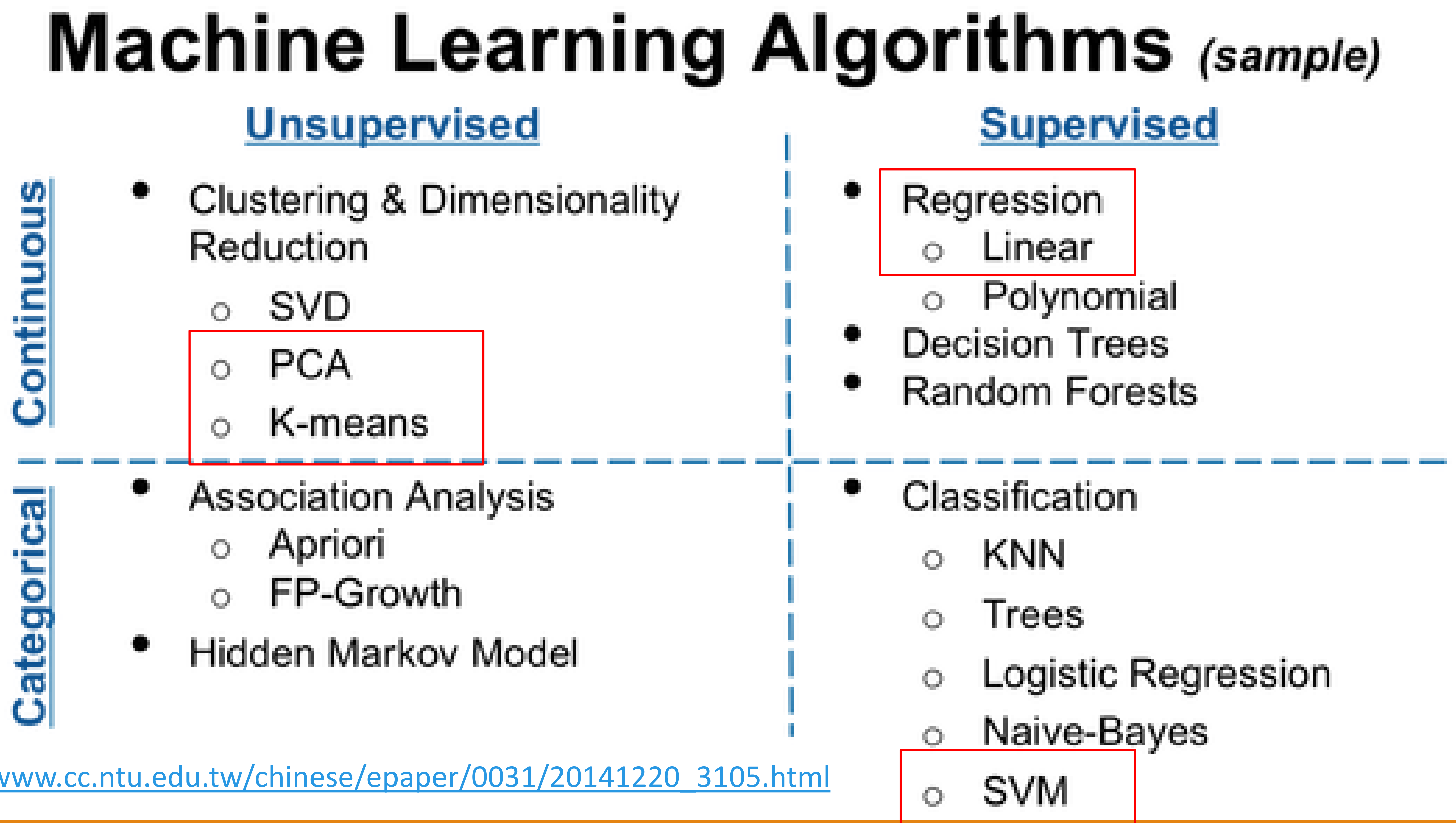
http://www.kdd.org/kdd2017/

機器學習是甚麼？


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
NO
YES
predicting a category
do you have labeled data
YES
NO
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
NOT WORKING
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering GMM
MiniBatch KMeans
<10K samples
MeanShift VBGMM
regression
<100K samples
SGD Regressor
few features should be important
ElasticNet Lasso
RidgeRegression SVR (kernel='linear')
SVR(kernel='rbf') EnsembleRegressors
dimensionality reduction
Randomized PCA
<10K samples
Isomap Spectral Embedding
LLE
kernel approximation



http://file.scirp.org/Html/6-9601216_35397.htm

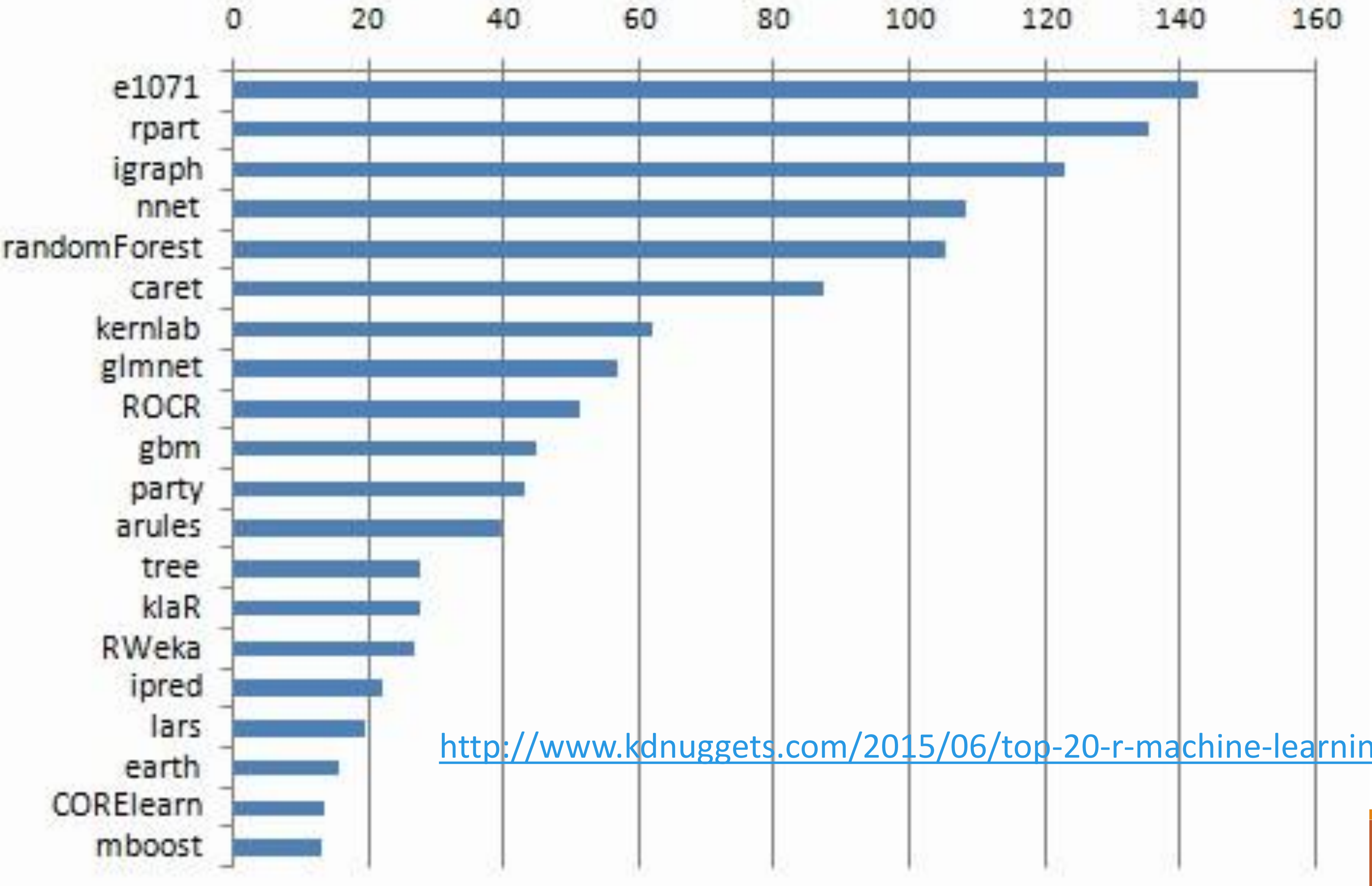


Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

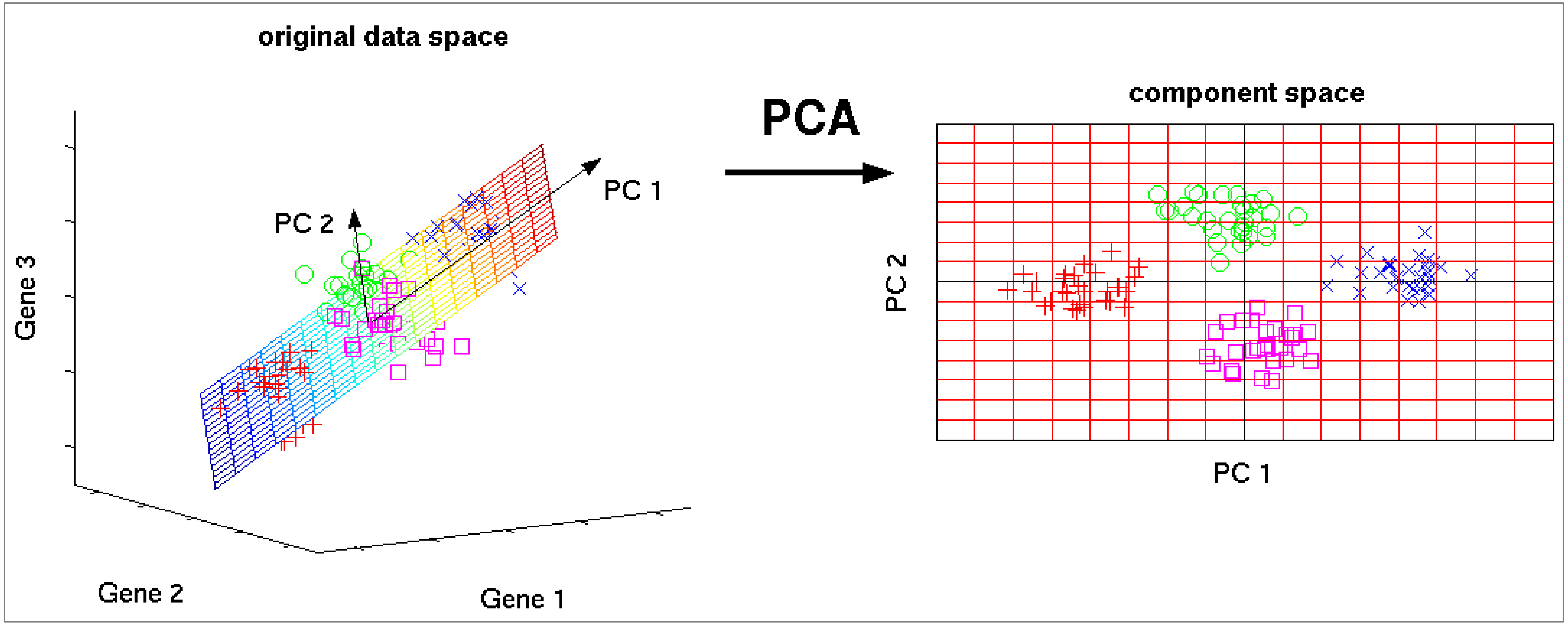
http://www.cc.ntu.edu.tw/chinese/epaper/0031/20141220_3105.html

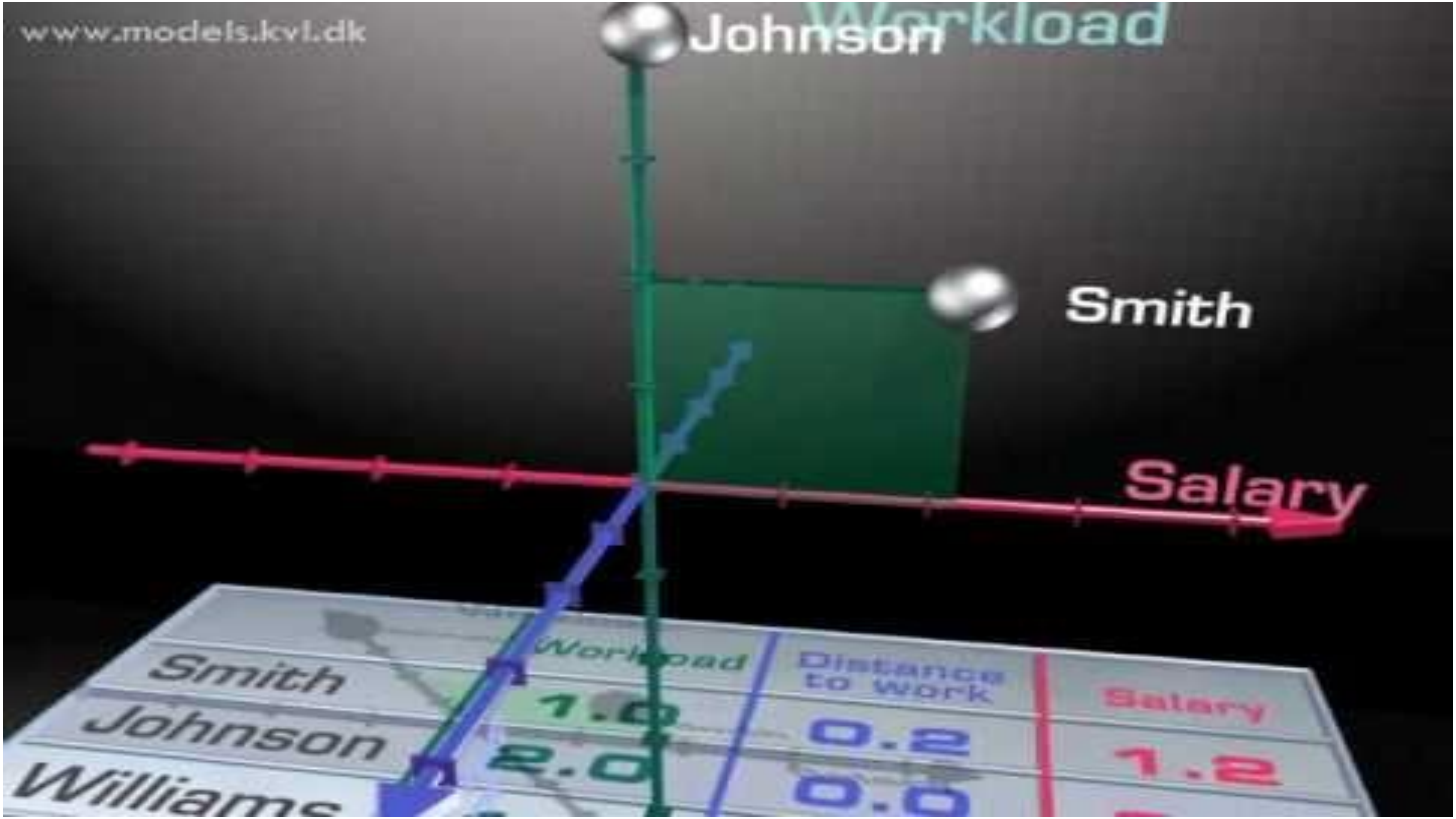
Top 20 R Machine Learning packages, by Downloads (000) from CRAN



http://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html

Principal Components Analysis





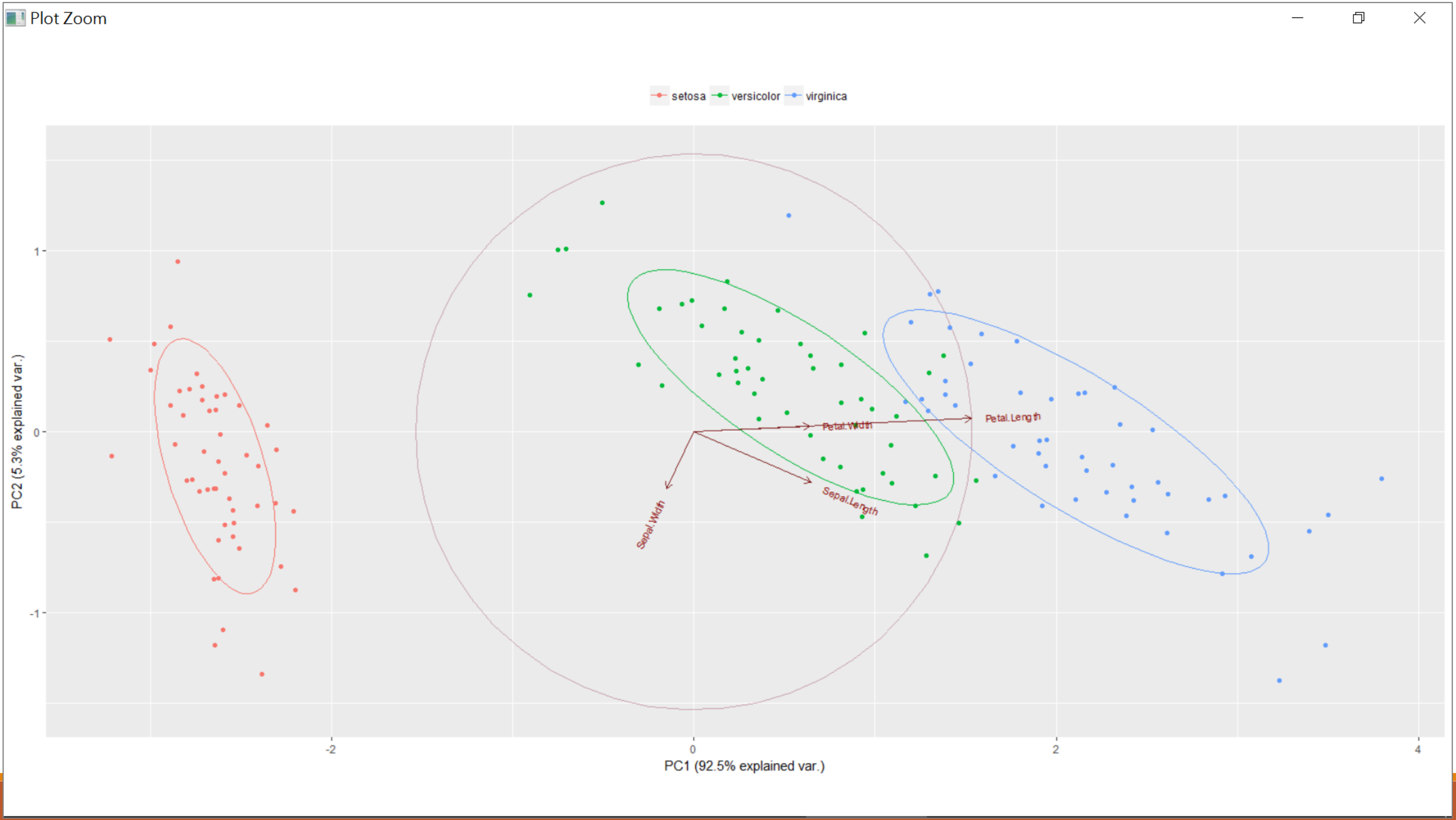
	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	[illegible]
Williams			

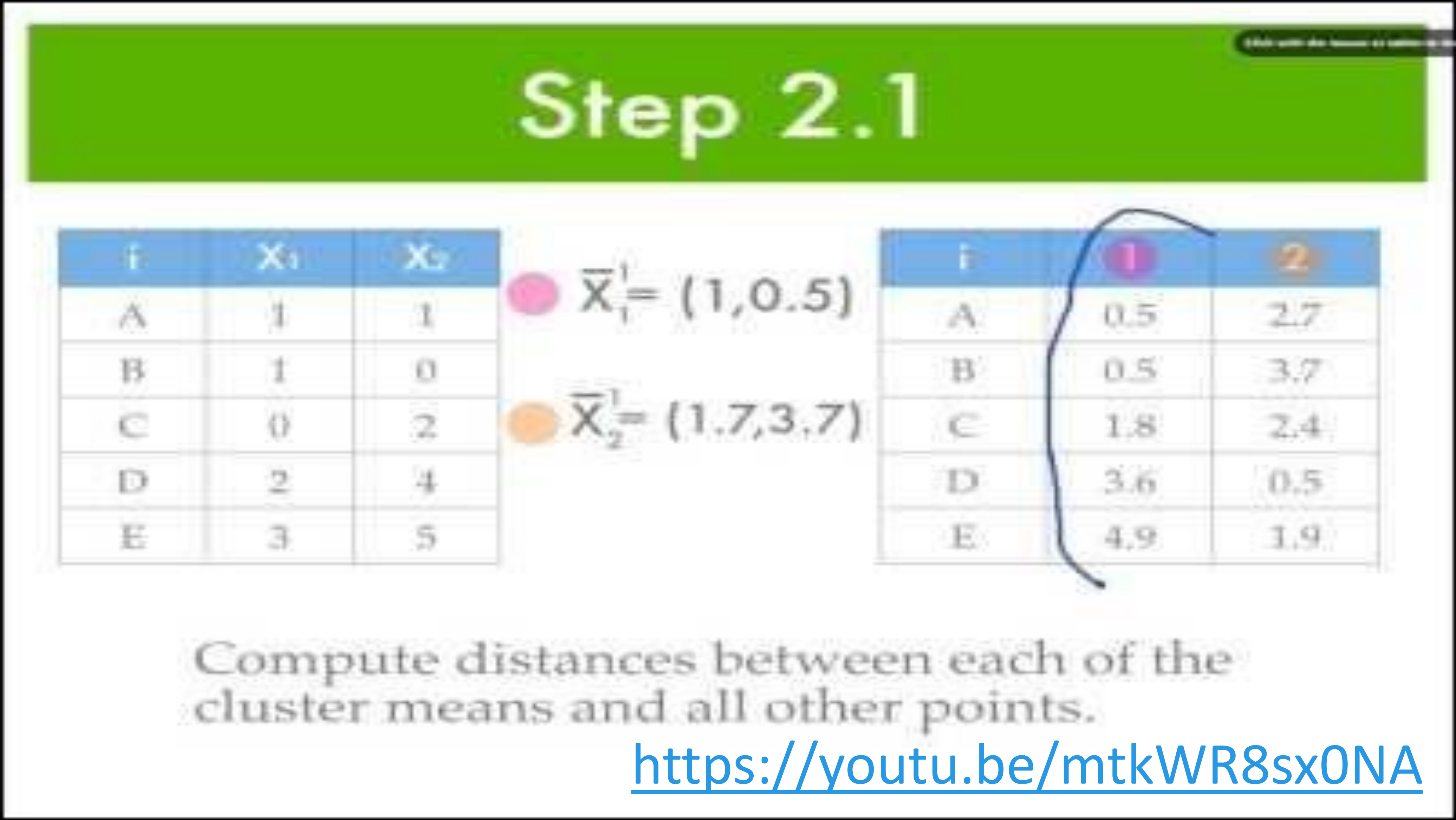
https://youtu.be/4pnQd6jnCWk

套件安裝

https://github.com/vqv/ggbiplot

```
library(devtools)
install_github("ggbiplot", "vqv")
library(scales)
library(grid)
library(ggbiplot)
```


Plot Zoom
setosa
versicolor
virginica
1
0
-1
-2
0
2
4
PC1 (92.5% explained var.)
PC2 (5.3% explained var.)
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width



Step 2.1

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

● $\bar{X}_1^1 = (1,0.5)$

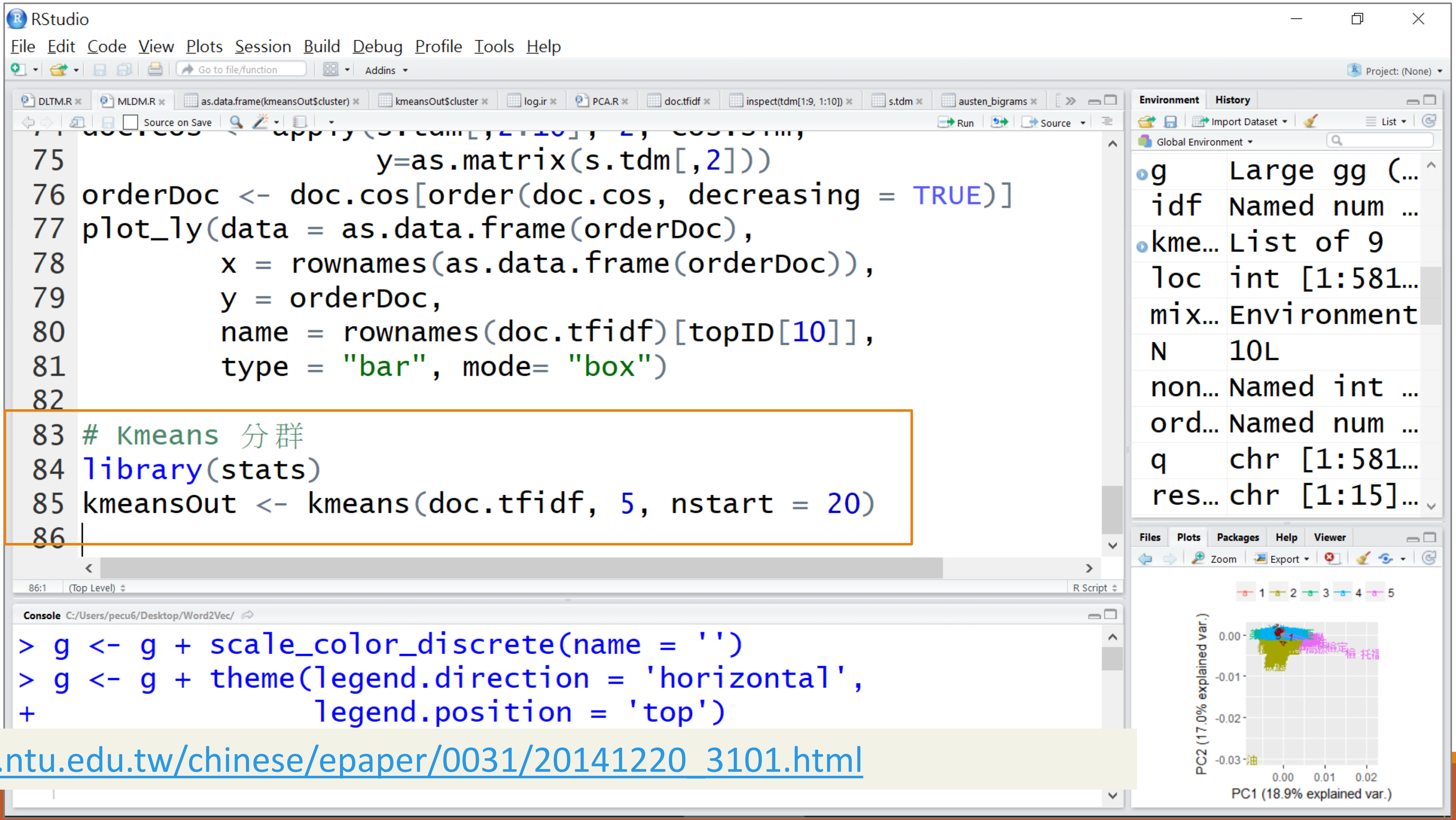
● $\bar{X}_2^1 = (1.7,3.7)$

i	1	2
A	0.5	2.7
B	0.5	3.7
C	1.8	2.4
D	3.6	0.5
E	4.9	1.9

Compute distances between each of the cluster means and all other points.

https://youtu.be/mtkWR8sx0NA

以 TF-IDF 進行 K-Means 分群



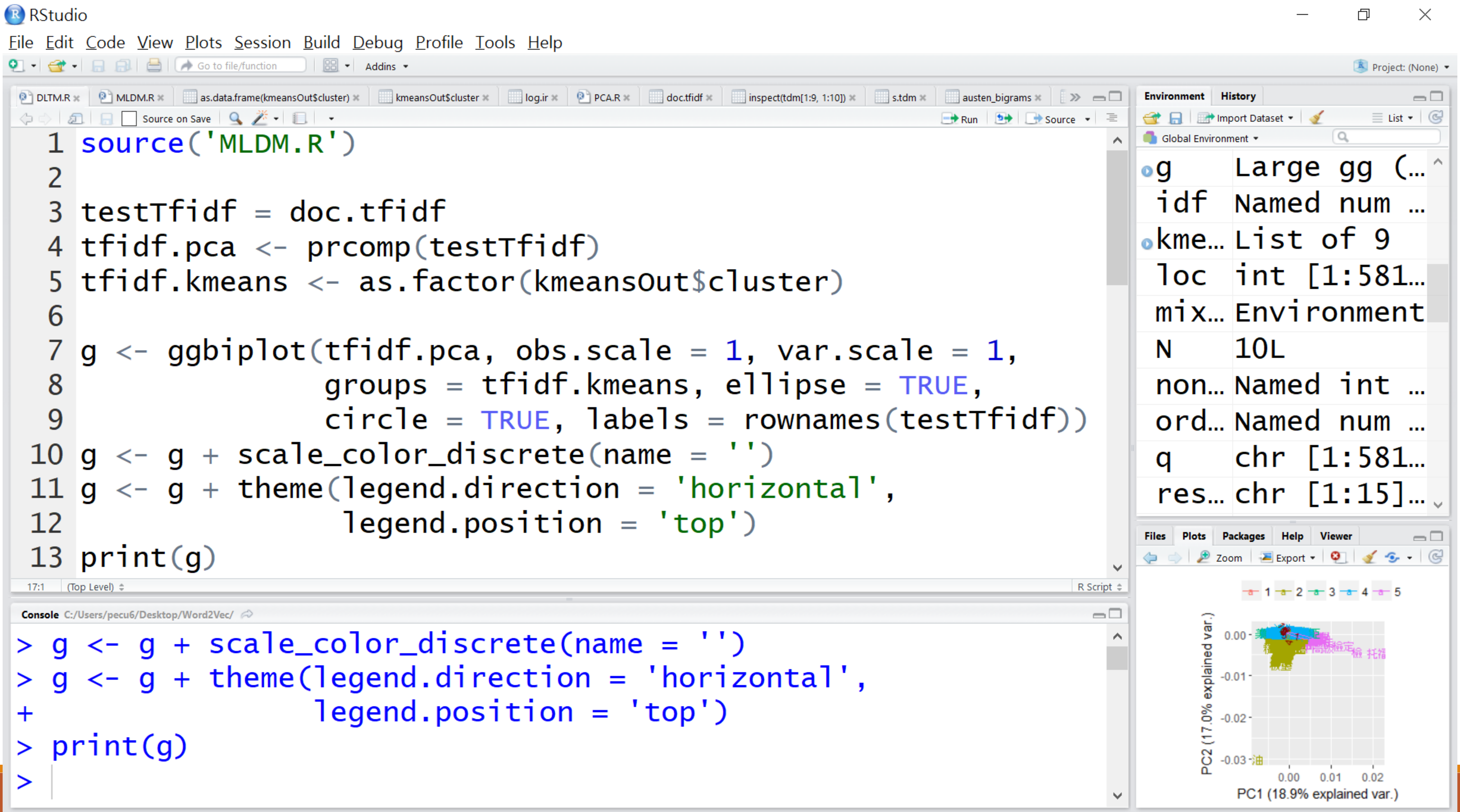
```r
doc.cos <- apply(s.tdm[,2:10], 2, cosSim,
                 y=as.matrix(s.tdm[,2]))
orderDoc <- doc.cos[order(doc.cos, decreasing = TRUE)]
plot_ly(data = as.data.frame(orderDoc),
        x = rownames(as.data.frame(orderDoc)),
        y = orderDoc,
        name = rownames(doc.tfidf)[topID[10]],
        type = "bar", mode= "box")

# Kmeans 分群
library(stats)
kmeansOut <- kmeans(doc.tfidf, 5, nstart = 20)

```

```
g    Large gg (…
idf  Named num …
kme… List of 9
loc  int [1:581…
mix… Environment
N    10L
non… Named int …
ord… Named num …
q    chr [1:581…
res… chr [1:15]…
```

Console C:/Users/pecu6/Desktop/Word2Vec/

```
> g <- g + scale_color_discrete(name = '')
> g <- g + theme(legend.direction = 'horizontal',
+                legend.position = 'top')
```

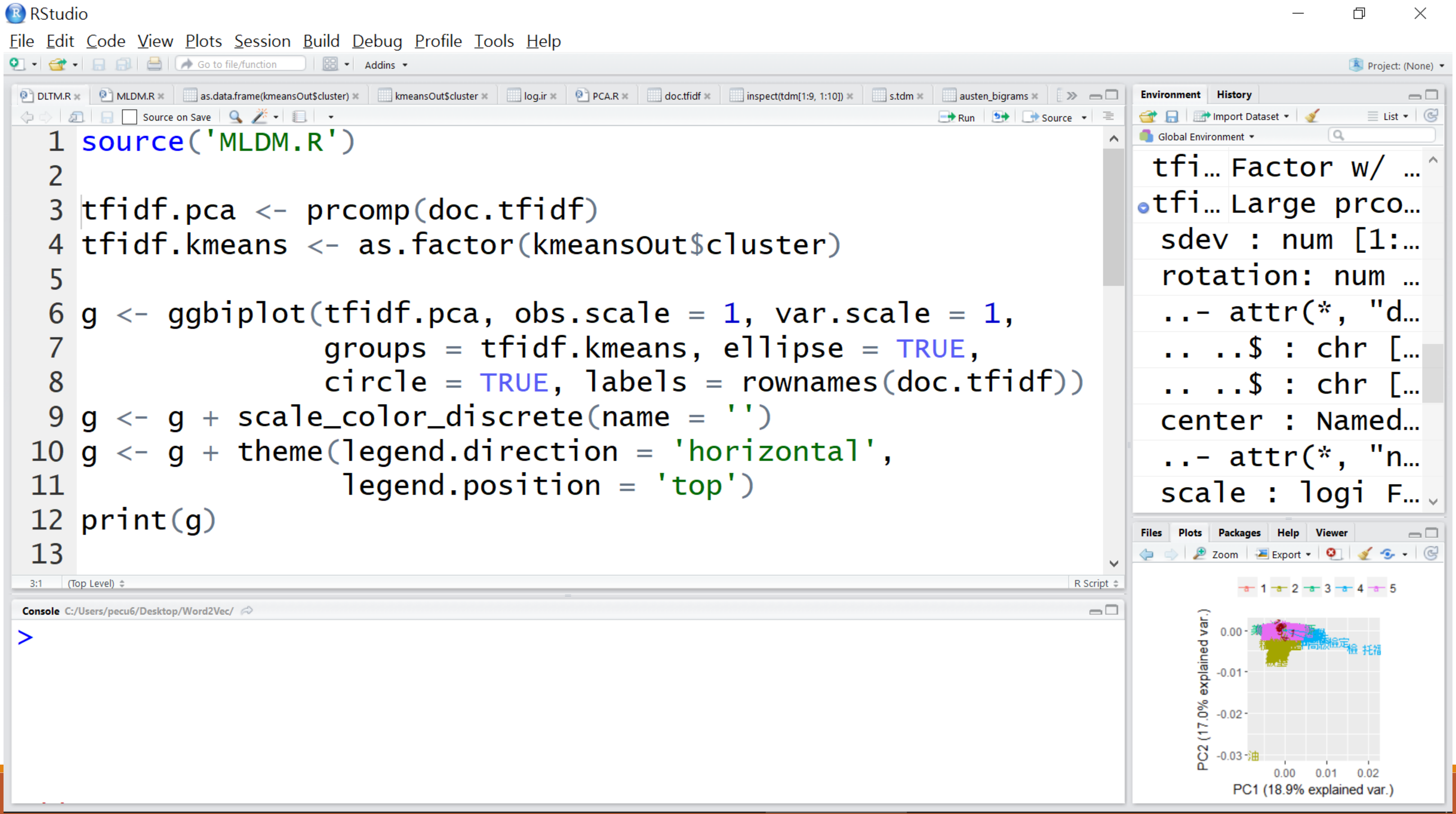
http://www.cc.ntu.edu.tw/chinese/epaper/0031/20141220_3101.html



```r
source('MLDM.R')

testTfidf = doc.tfidf
tfidf.pca <- prcomp(testTfidf)
tfidf.kmeans <- as.factor(kmeansOut$cluster)

g <- ggbiplot(tfidf.pca, obs.scale = 1, var.scale = 1,
              groups = tfidf.kmeans, ellipse = TRUE,
              circle = TRUE, labels = rownames(testTfidf))
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```

```
g       Large gg (…
idf     Named num …
kme…    List of 9
loc     int [1:581…
mix…    Environment
N       10L
non…    Named int …
ord…    Named num …
q       chr [1:581…
res…    chr [1:15]…
```

```
> g <- g + scale_color_discrete(name = '')
> g <- g + theme(legend.direction = 'horizontal',
+                legend.position = 'top')
> print(g)
>
```

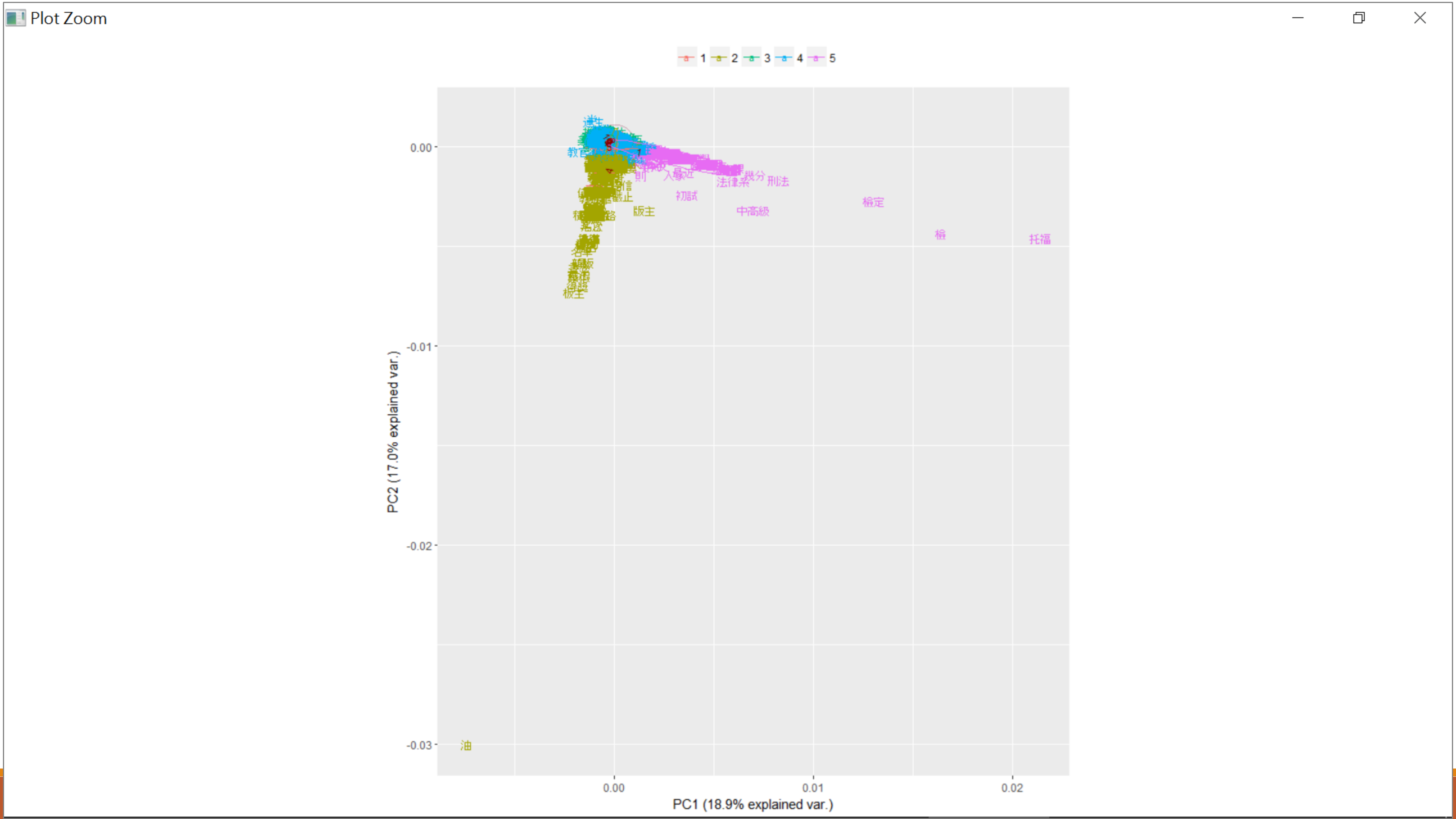


```r
source('MLDM.R')

tfidf.pca <- prcomp(doc.tfidf)
tfidf.kmeans <- as.factor(kmeansOut$cluster)

g <- ggbiplot(tfidf.pca, obs.scale = 1, var.scale = 1,
              groups = tfidf.kmeans, ellipse = TRUE,
              circle = TRUE, labels = rownames(doc.tfidf))
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)

```

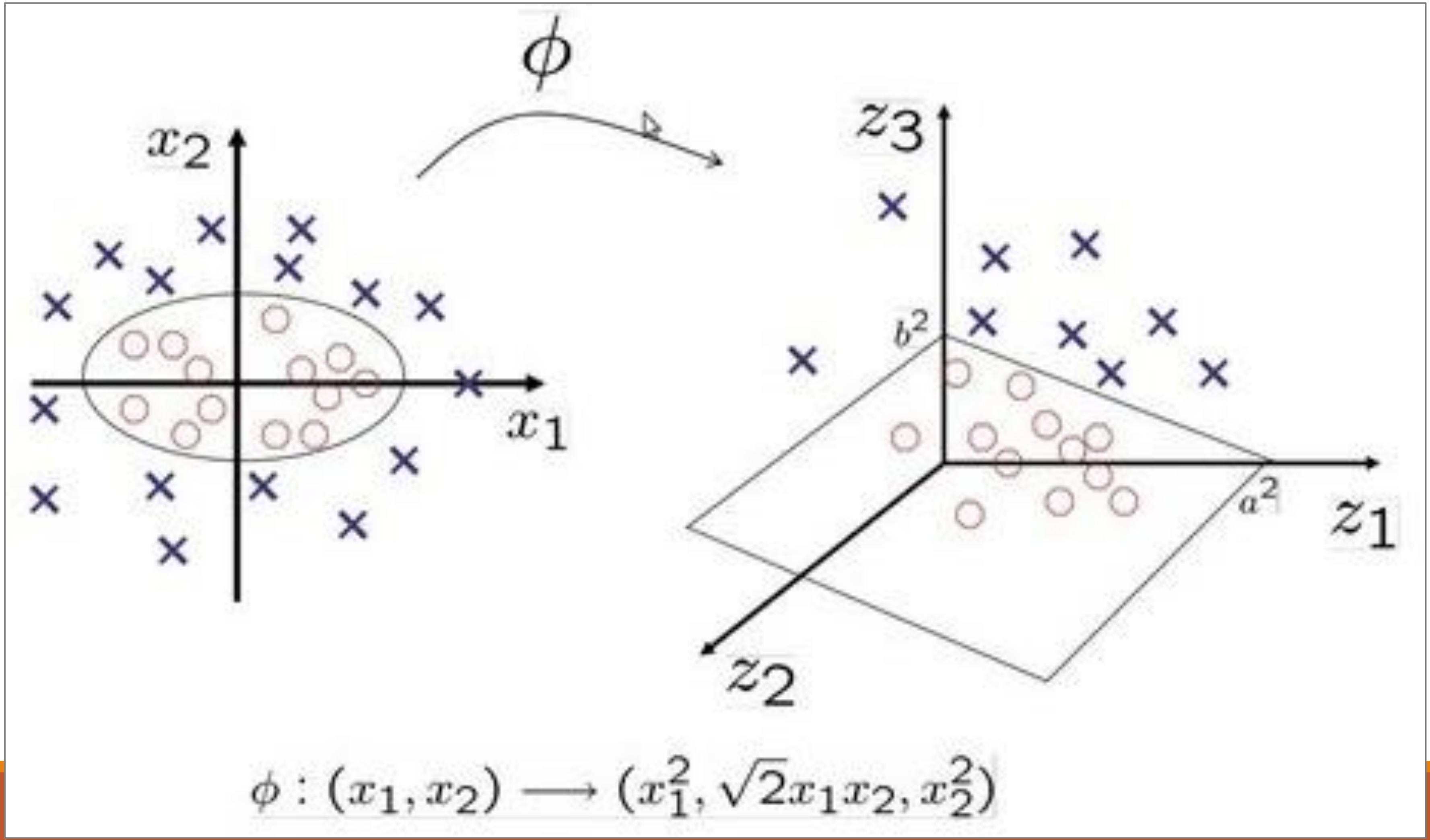
```
tfi… Factor w/ …
tfi… Large prco…
 sdev : num [1:…
 rotation: num …
 ..- attr(*, "d…
 .. ..$ : chr […
 .. ..$ : chr […
 center : Named…
 ..- attr(*, "n…
 scale : logi F…
```

Console C:/Users/pecu6/Desktop/Word2Vec/

```
>
```

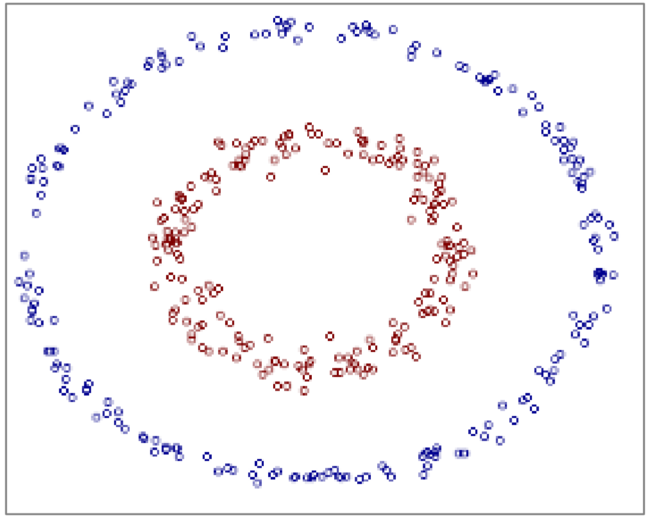

Plot Zoom
1
2
3
4
5
遊生
教官
0.00
-0.01
-0.02
-0.03
則
入學
最近
法律系
幾分
刑法
信
截止
初試
版主
路
中高級
檢定
檢
托福
板主
油
0.00
0.01
0.02
PC2 (17.0% explained var.)
PC1 (18.9% explained var.)

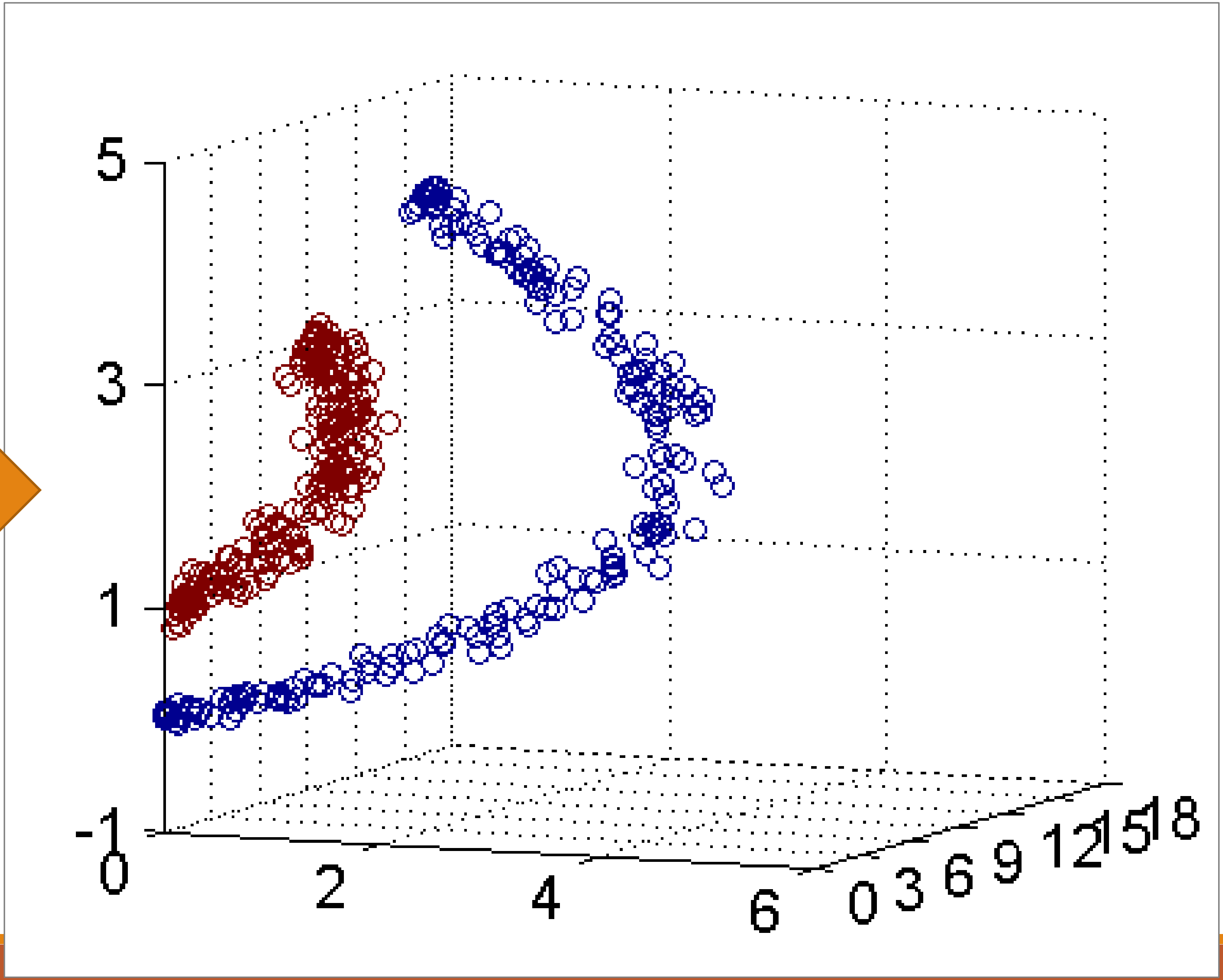
Support Vector Machine



$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

https://drive.google.com/file/d/0BxNbusWG4rLgNkRqZGpUdW9jV2c/view?usp=sharing

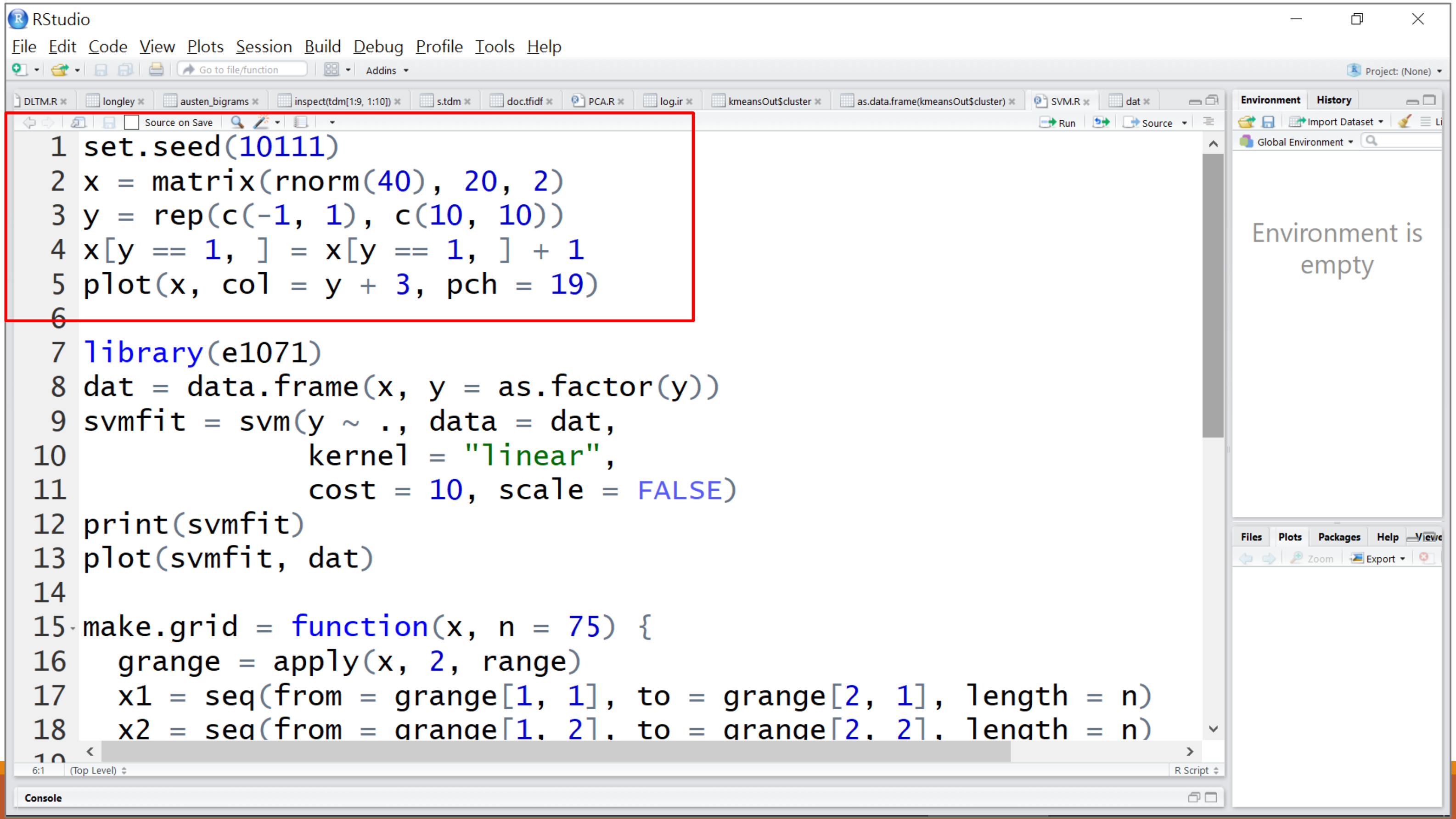




Support Vector Regression with R

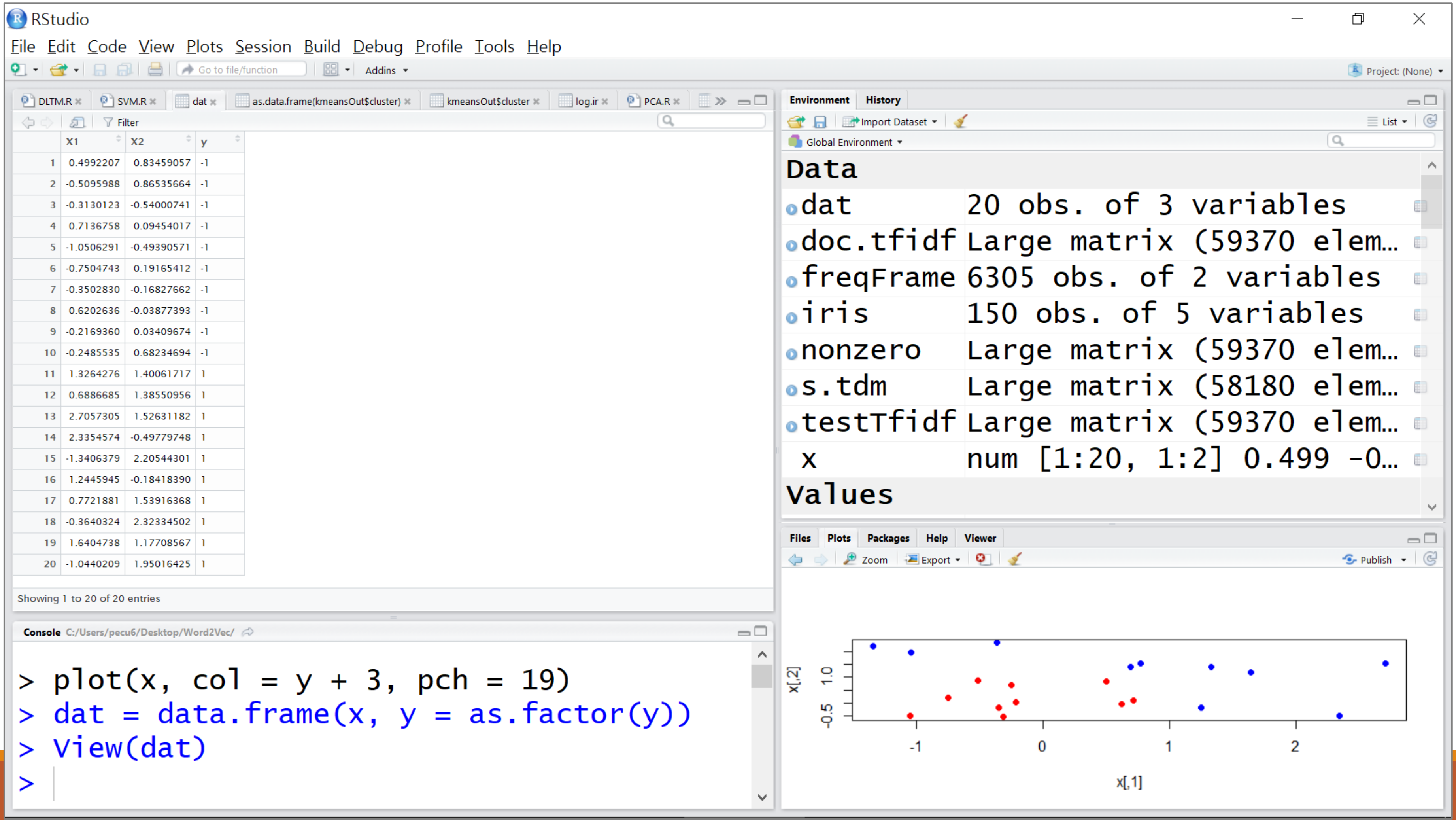
https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/ch9.html

Linear 資料準備



```r
set.seed(10111)
x = matrix(rnorm(40), 20, 2)
y = rep(c(-1, 1), c(10, 10))
x[y == 1, ] = x[y == 1, ] + 1
plot(x, col = y + 3, pch = 19)

library(e1071)
dat = data.frame(x, y = as.factor(y))
svmfit = svm(y ~ ., data = dat,
             kernel = "linear",
             cost = 10, scale = FALSE)
print(svmfit)
plot(svmfit, dat)

make.grid = function(x, n = 75) {
  grange = apply(x, 2, range)
  x1 = seq(from = grange[1, 1], to = grange[2, 1], length = n)
  x2 = seq(from = grange[1, 2], to = grange[2, 2], length = n)
```

使用套件：library(e1071)



	X1	X2	y
1	0.4992207	0.83459057	-1
2	-0.5095988	0.86535664	-1
3	-0.3130123	-0.54000741	-1
4	0.7136758	0.09454017	-1
5	-1.0506291	-0.49390571	-1
6	-0.7504743	0.19165412	-1
7	-0.3502830	-0.16827662	-1
8	0.6202636	-0.03877393	-1
9	-0.2169360	0.03409674	-1
10	-0.2485535	0.68234694	-1
11	1.3264276	1.40061717	1
12	0.6886685	1.38550956	1
13	2.7057305	1.52631182	1
14	2.3354574	-0.49779748	1
15	-1.3406379	2.20544301	1
16	1.2445945	-0.18418390	1
17	0.7721881	1.53916368	1
18	-0.3640324	2.32334502	1
19	1.6404738	1.17708567	1
20	-1.0440209	1.95016425	1

Showing 1 to 20 of 20 entries

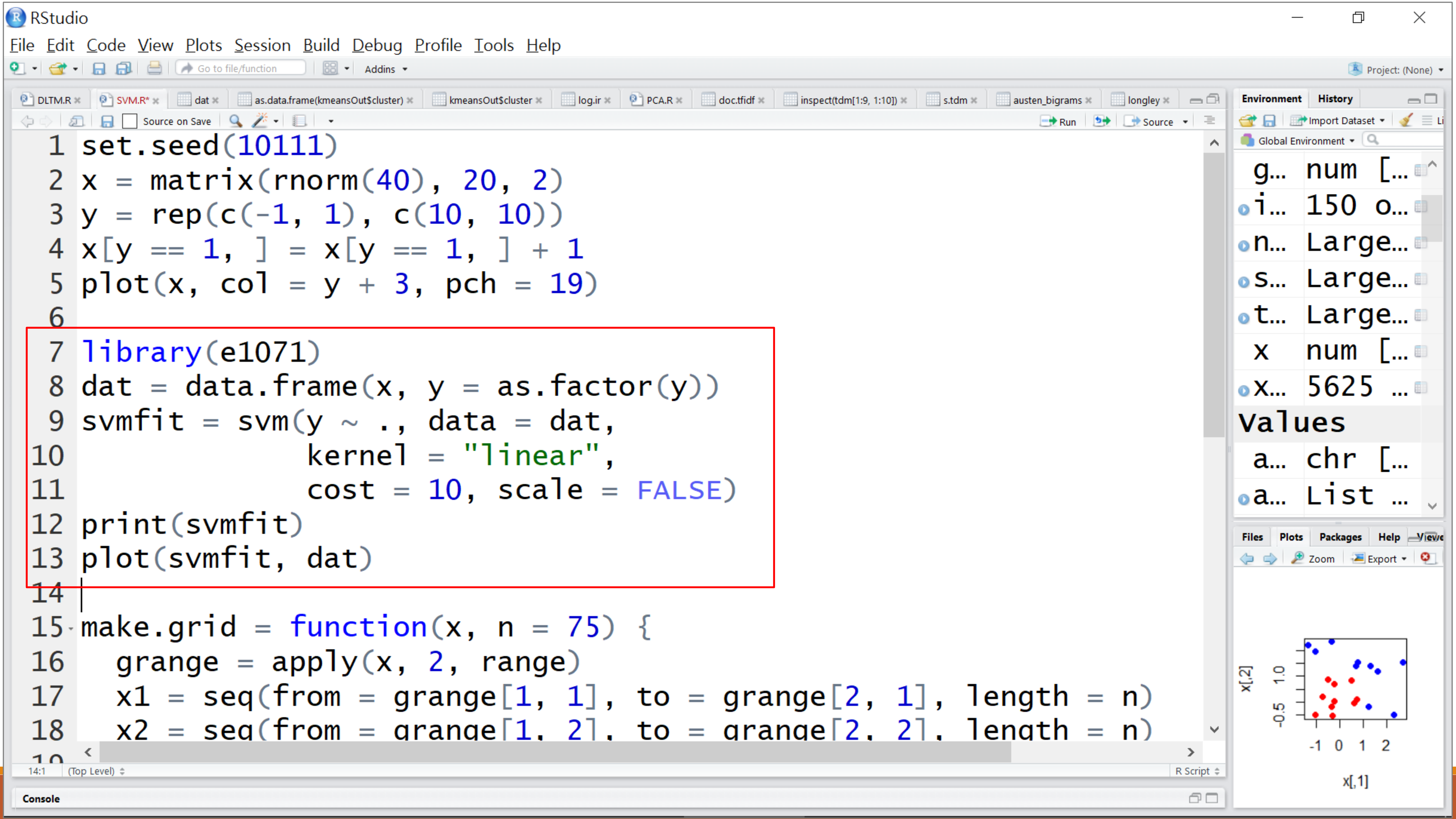
Data

dat	20 obs. of 3 variables
doc.tfidf	Large matrix (59370 elem…
freqFrame	6305 obs. of 2 variables
iris	150 obs. of 5 variables
nonzero	Large matrix (59370 elem…
s.tdm	Large matrix (58180 elem…
testTfidf	Large matrix (59370 elem…
x	num [1:20, 1:2] 0.499 -0…

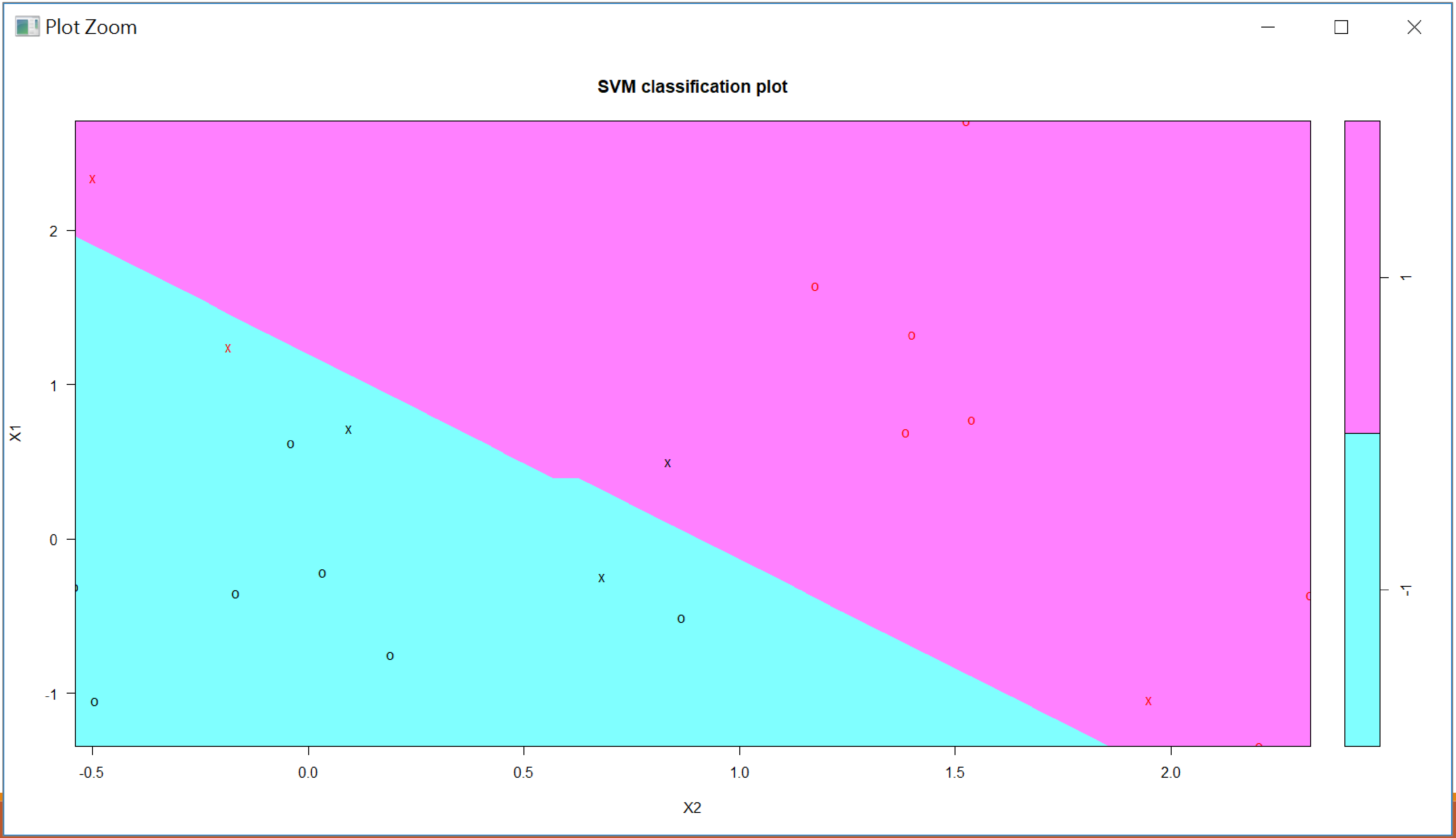
Values

Console C:/Users/pecu6/Desktop/Word2Vec/

```
> plot(x, col = y + 3, pch = 19)
> dat = data.frame(x, y = as.factor(y))
> View(dat)
> 
```

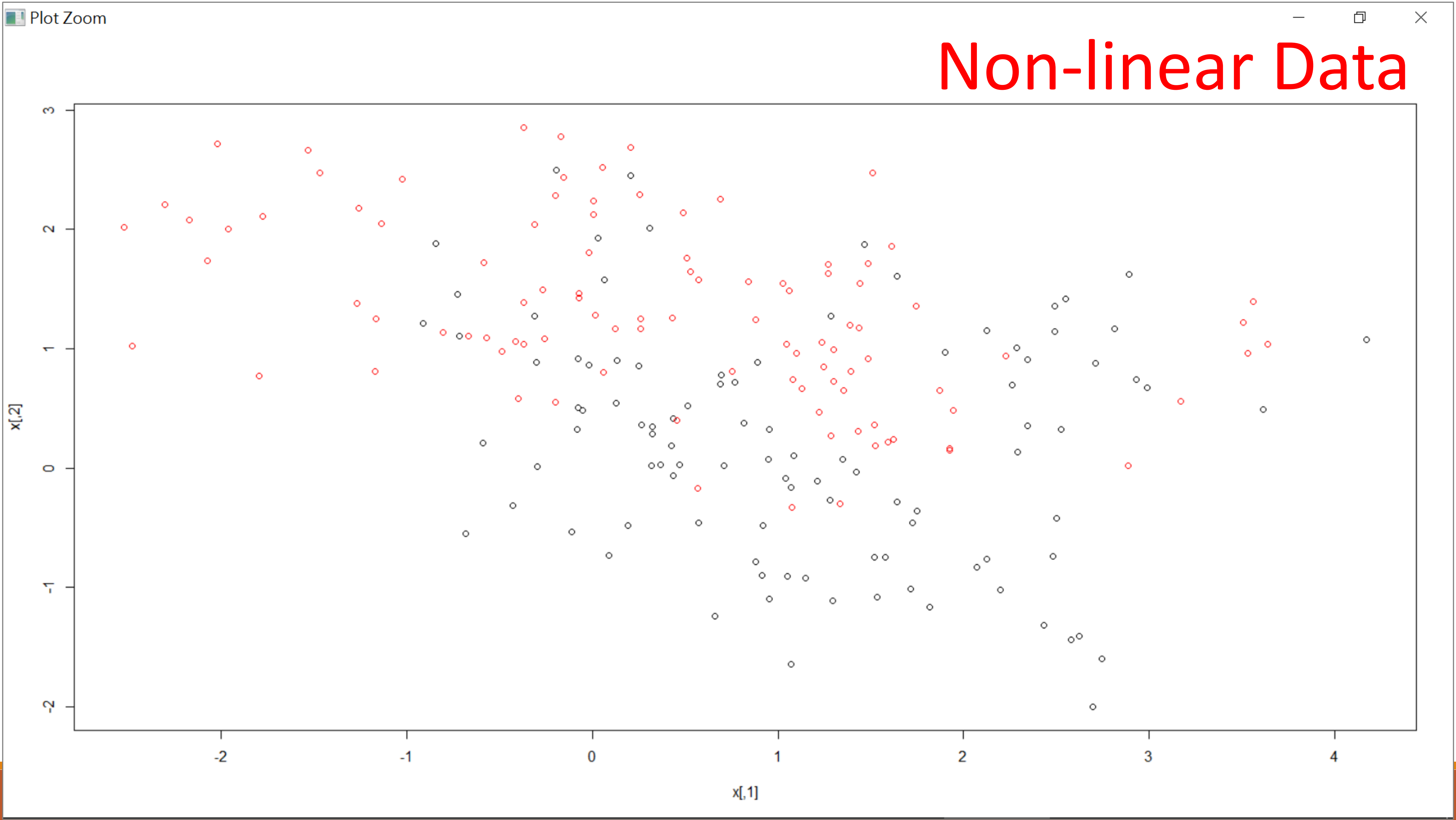


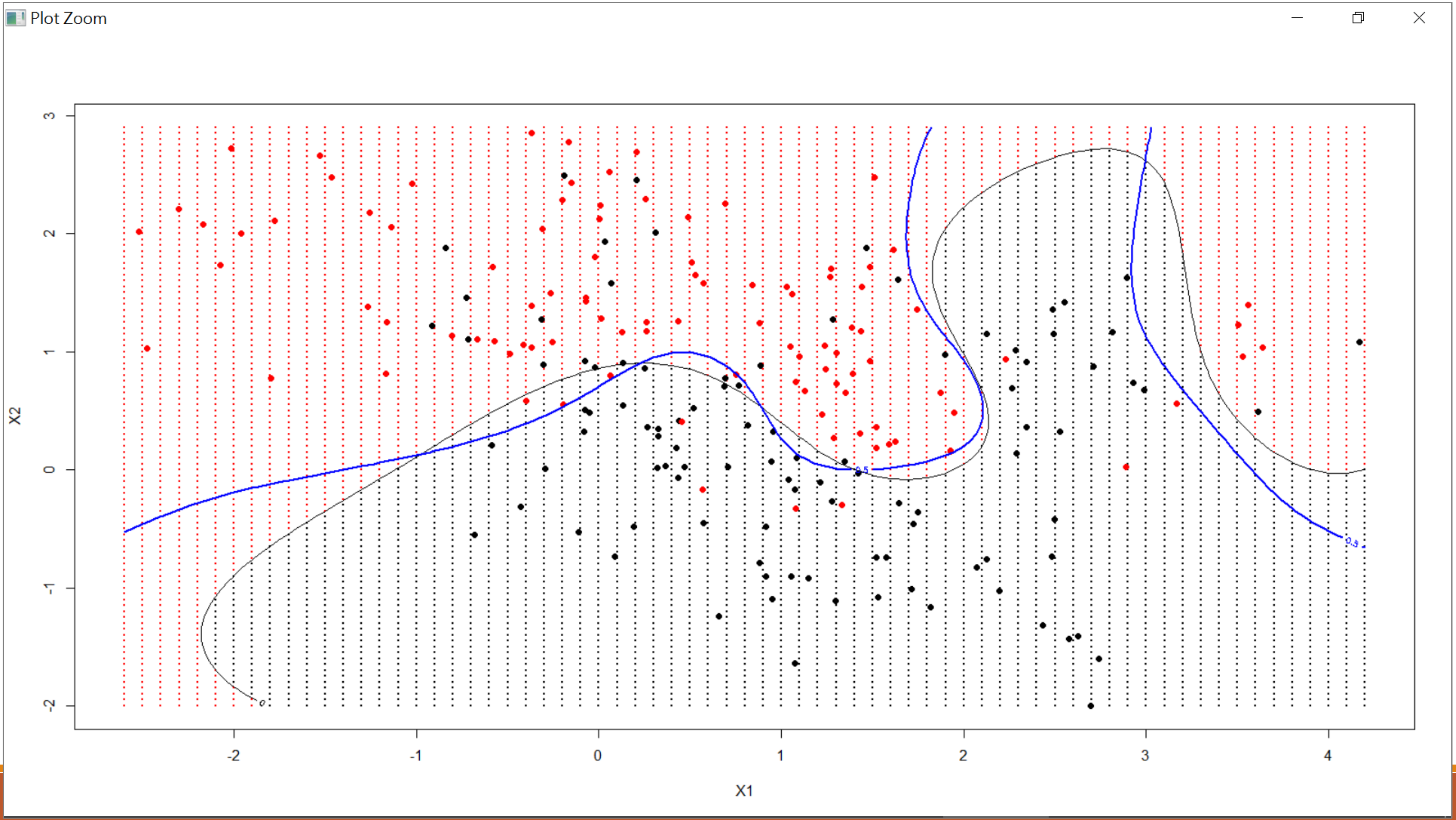
```
set.seed(10111)
x = matrix(rnorm(40), 20, 2)
y = rep(c(-1, 1), c(10, 10))
x[y == 1, ] = x[y == 1, ] + 1
plot(x, col = y + 3, pch = 19)

library(e1071)
dat = data.frame(x, y = as.factor(y))
svmfit = svm(y ~ ., data = dat,
             kernel = "linear",
             cost = 10, scale = FALSE)
print(svmfit)
plot(svmfit, dat)

make.grid = function(x, n = 75) {
  grange = apply(x, 2, range)
  x1 = seq(from = grange[1, 1], to = grange[2, 1], length = n)
  x2 = seq(from = grange[1, 2], to = grange[2, 2], length = n)
```

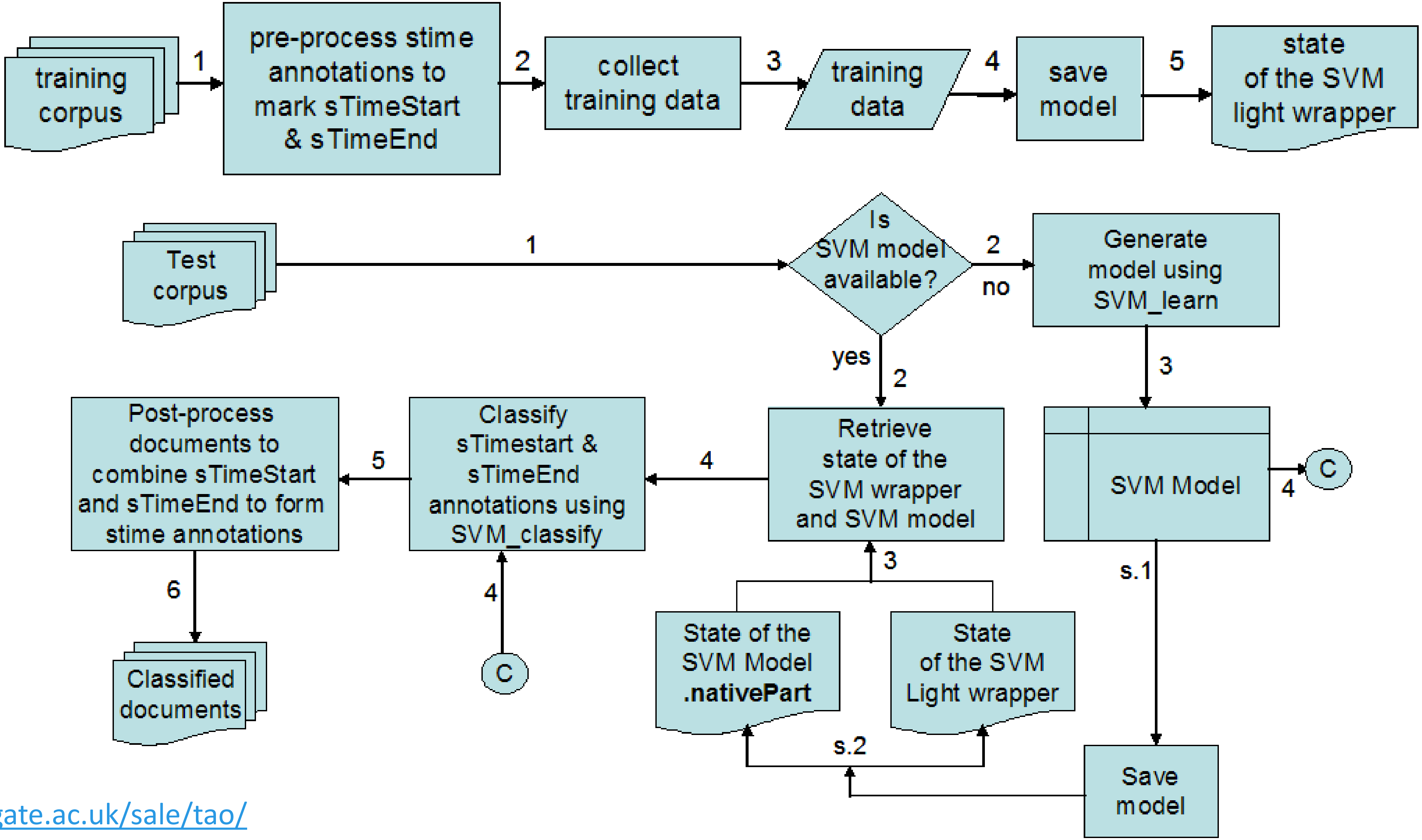


Plot Zoom
SVM classification plot
X1
2
1
0
-1
X2
-0.5
0.0
0.5
1.0
1.5
2.0
1
-1

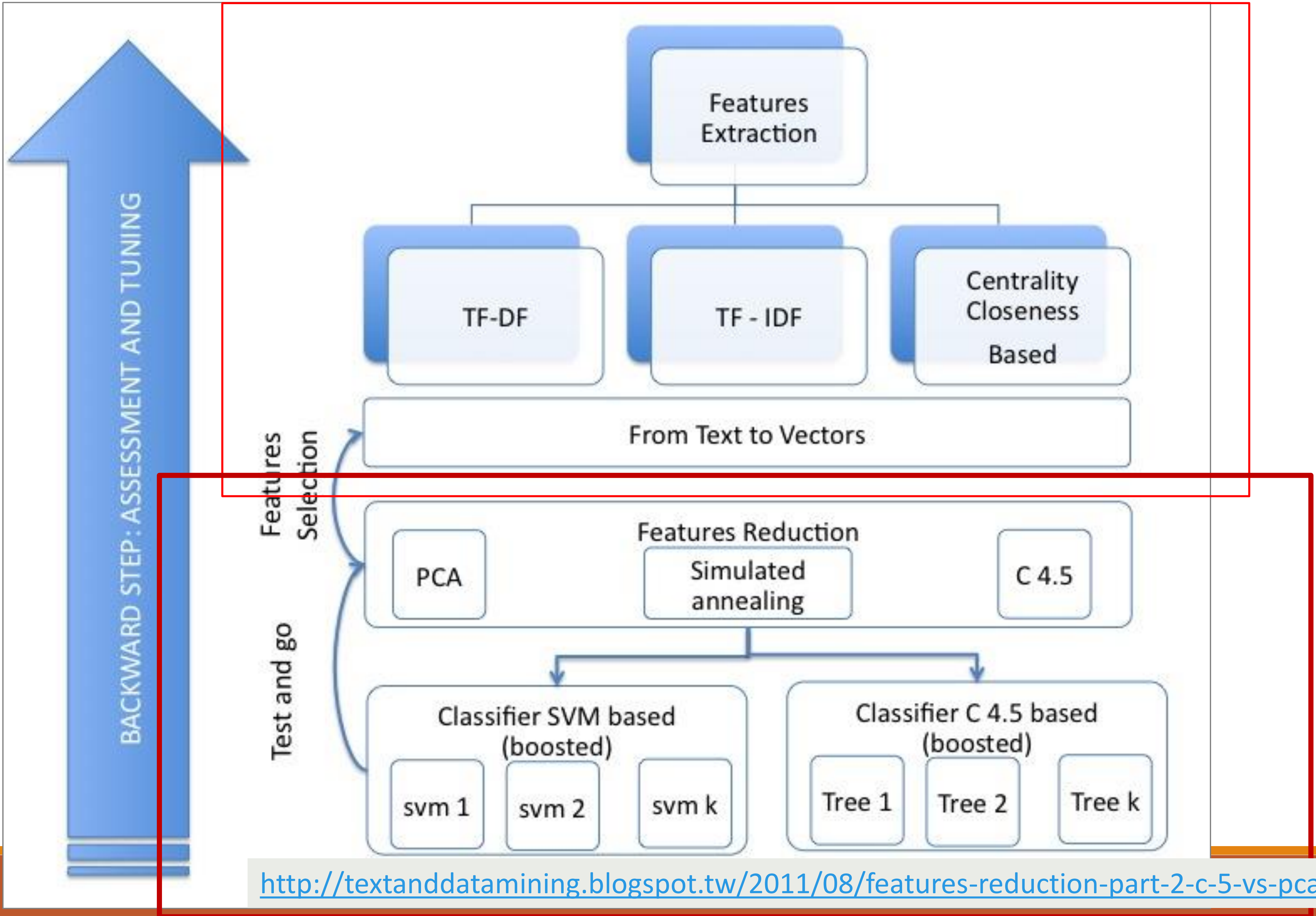
Non-linear Data




Plot Zoom
X2
X1
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
0
0.5
0.5


training corpus
1
pre-process stime annotations to mark sTimeStart & sTimeEnd
2
collect training data
3
training data
4
save model
5
state of the SVM light wrapper
Test corpus
1
Is SVM model available?
2
no
Generate model using SVM_learn
yes
2
3
Post-process documents to combine sTimeStart and sTimeEnd to form stime annotations
5
Classify sTimestart & sTimeEnd annotations using SVM_classify
4
Retrieve state of the SVM wrapper and SVM model
SVM Model
4
C
6
4
3
s.1
Classified documents
C
State of the SVM Model .nativePart
State of the SVM Light wrapper
s.2
Save model

https://gate.ac.uk/sale/tao/



http://textanddatamining.blogspot.tw/2011/08/features-reduction-part-2-c-5-vs-pca.html