Big Data Statistics

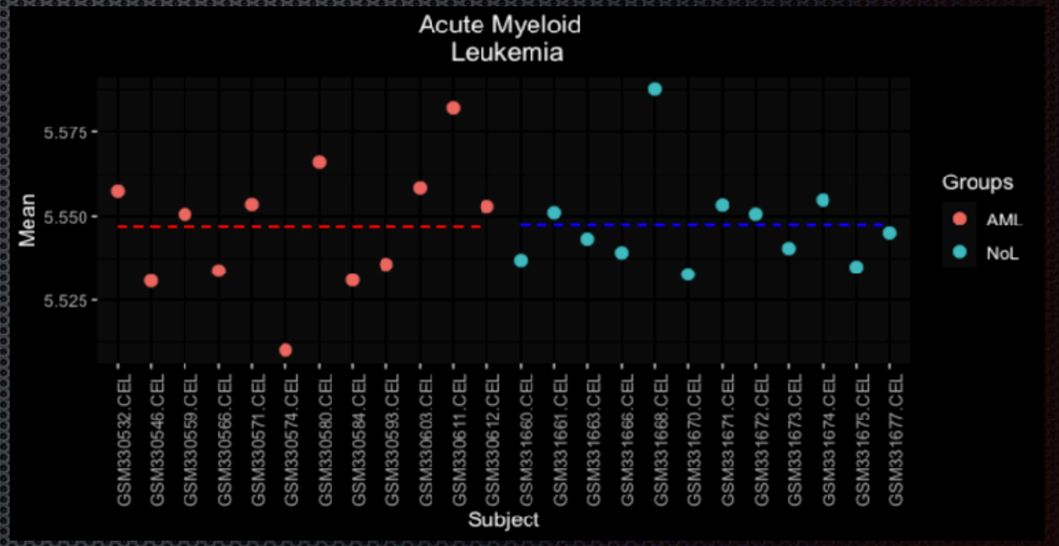
Efthymios Ioannis Kavour

Table of Content

Dimension Reduction and Big Data Statistics

Working with big data is a challenging topic. Conventional Statistics either will not work or it will consume a lot of resources (inc. time).

20172 x 24

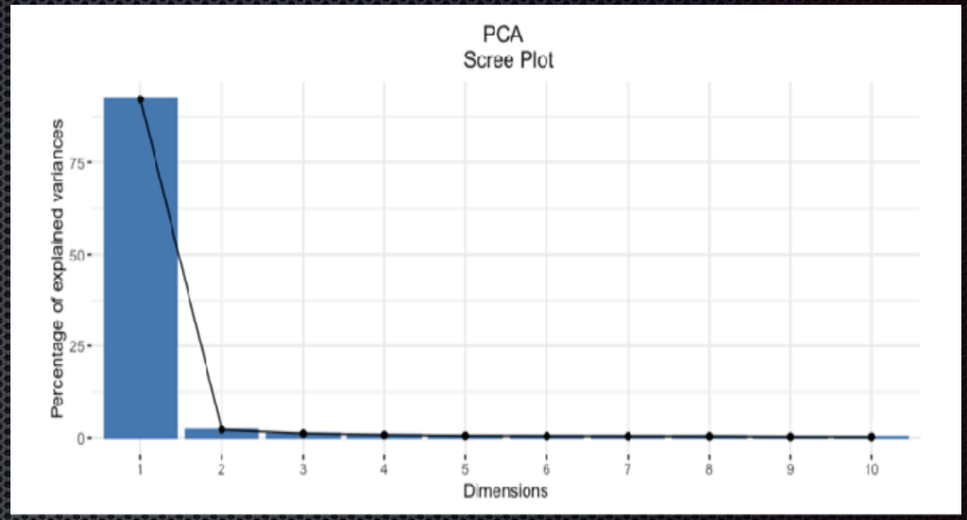


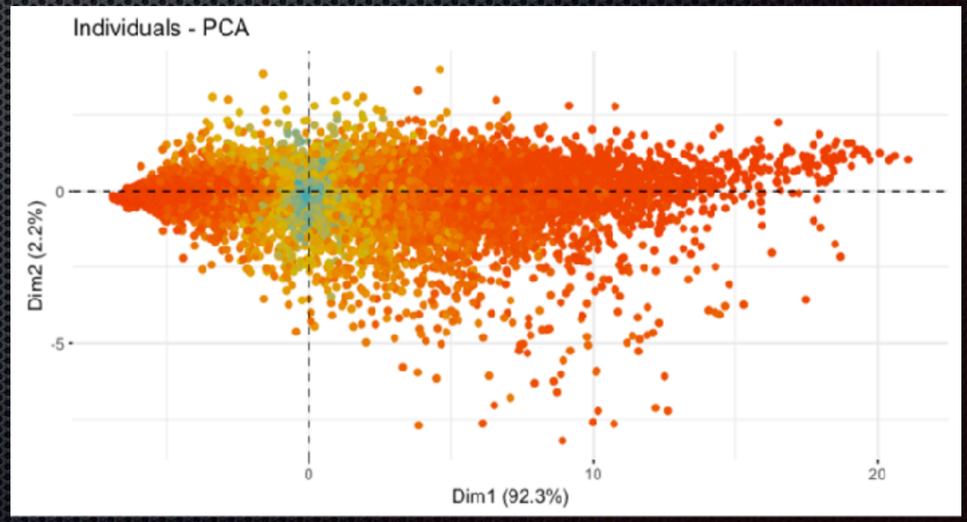
24 samples =
24 dimensions =
24 problems...

Dimension Reduction and Big Data Statistics

Principal Component Analysis is our solution.

92.3 % of data variance are explained by the first PC whereas 2.2 % of by the second PC.





Dimension Reduction and Big Data Statistics

But what about Hypothesis testing?

One can apply multiple hypothesis testing

FWER	92
FDR	953
pFDR	1687

Multiple test hypothesis need some adjustments.
q_value is a solution
q_value is the minimum FDR and allow us simultaneous check of multiple results.

But the significant level is as usual right?

Unfortunately not, action need to be taken!

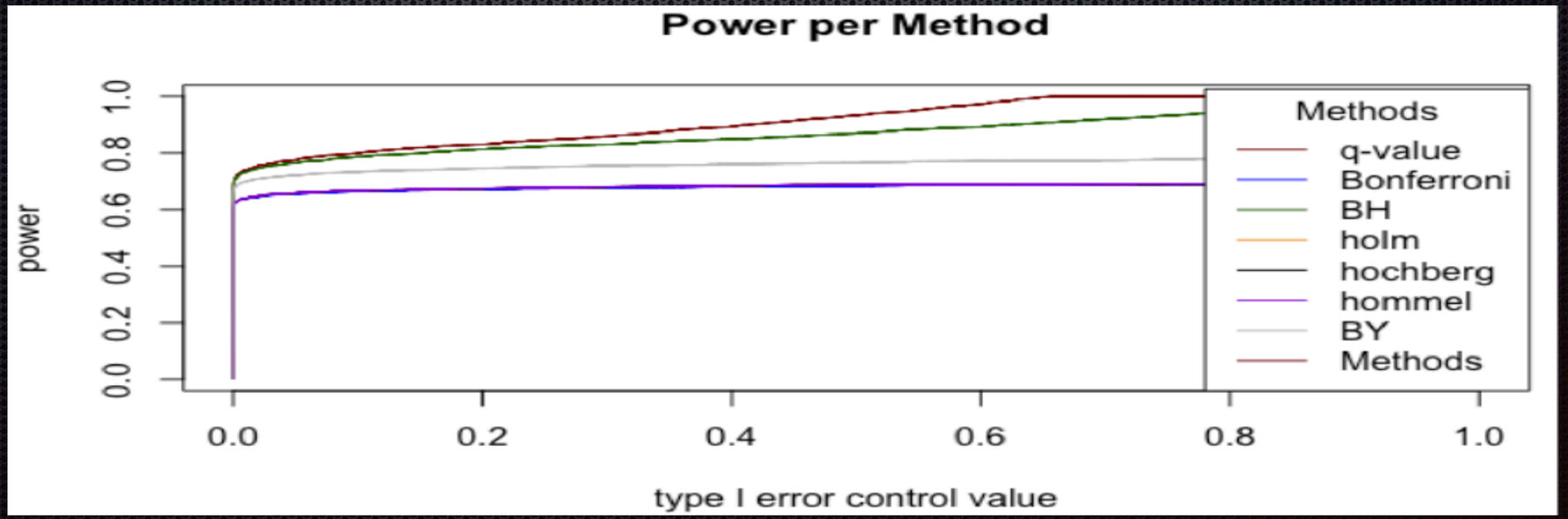
- Benjamini Hochberg
- Bonferonni
- Holm
- Hochberg

Dimension Reduction and Big Data Statistics

How to choose the right one?

In order to compare the different methods we use **power** a measurement of the proportion of Non-True that are decleared significant to the total True significant observations.

Method	Power
Bonferroni	0
BH	0.0130672
Holm	0
Hochberg	0
Hommel	0
BY	0.001148765
qvalue	0.02096496



Big Data Regression

Regression analysis is a powerful statistical technique used to understand and model the relationship between variables. In the context of big data, regression analysis plays a crucial role in extracting insights and making predictions.

The lm + update way

The matrix $(X^TX)^{-1}X^TY$ way

Big Data Regression

The lm + update Way

```
Large data regression model: bigglm(fit, matrix_list[[1]])
Sample size =  2e+06
              Coef    (95%     CI)    SE p
(Intercept)  0.4120  0.4106  0.4134 7e-04 0
x1           0.6082  0.6068  0.6096 7e-04 0
x2          -1.5170 -1.5184 -1.5155 7e-04 0
x3           0.7838  0.7824  0.7852 7e-04 0
x4          -0.0457 -0.0471 -0.0443 7e-04 0
x5          -0.1837 -0.1851 -0.1823 7e-04 0
x6          -1.2416 -1.2430 -1.2402 7e-04 0
...
```

The Matrix Way

	y
V1	0.41198072
x1	0.60822588
x2	-1.51695343
x3	0.78378959
x4	-0.04570068
x5	-0.18371295
x6	-1.24157476

=

Big Data Regression

Airline dataset Model. → 7129270x29 => 3018442 x6

glm model using gamma family → for loop with update

```
Call:
glm(formula = ArrDelay ~ Month + DayOfWeek + Distance +
                                    DepDelay +
    DepTime, family = Gamma(link = "log"), data = matrix_nb[[1]
                                    ])

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.589e+00  2.082e-02 124.335  < 2e-16 ***
Month               NA         NA      NA       NA
DayOfWeek   -1.106e-02  2.508e-03  -4.411 1.03e-05 ***
Distance     6.557e-05  9.270e-06   7.073 1.55e-12 ***
DepDelay     1.945e-02  1.255e-04 154.923  < 2e-16 ***
DepTime      1.045e-05  1.153e-05   0.906    0.365
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 0.8240305)

    Null deviance: 47986  on 32999  degrees of freedom
Residual deviance: 24518  on 32995  degrees of freedom
AIC: 264350

Number of Fisher Scoring iterations: 8
```

Big Data Regression

There are other methods that can help us select features from big data. One of which is lasso.

λ	Var.Selected
0.05	12
0.1	12
0.2	12
0.3	12
0.4	12
0.5	12
1.0	10
1.5	8
2.0	7
2.5	5

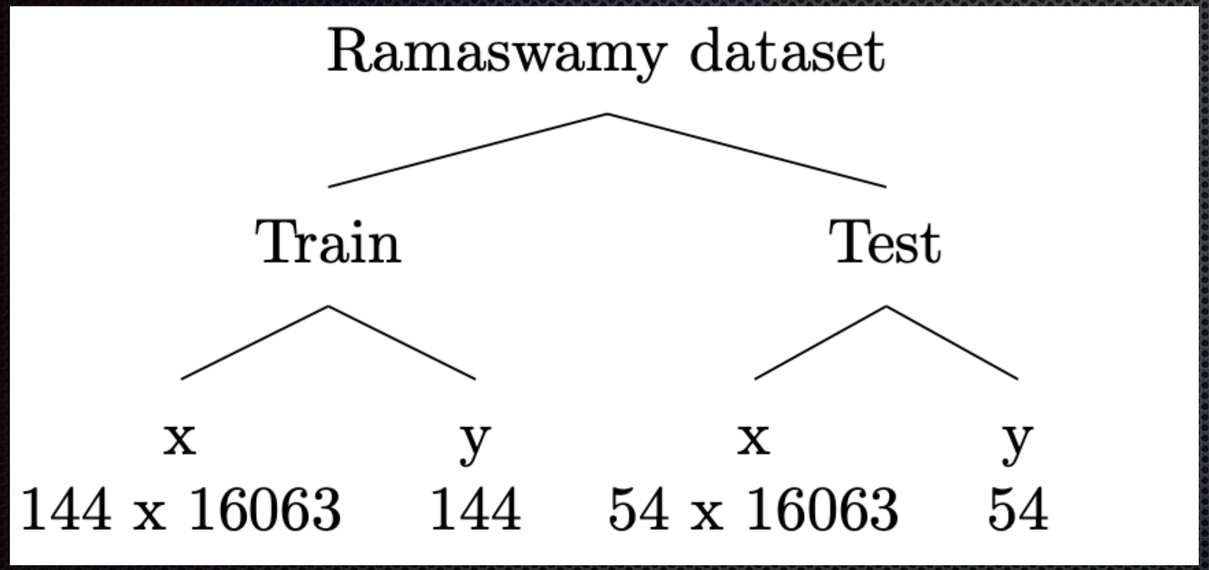
We used lasso to pick the select the most useful features.

GOAL:

$$minimize \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Machine Learning using Big Data

Classification ML on gene data

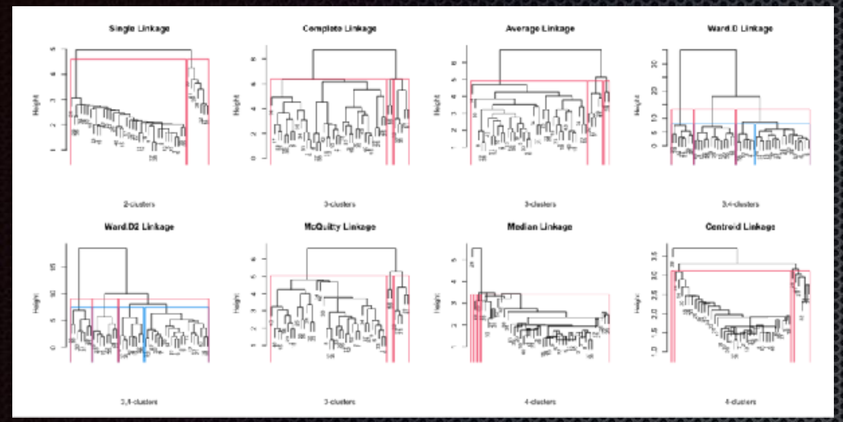


Classification Methods	
Method	Accuracy
Naive Bayes	42.59%
KNN	42.59%
Decision Trees	38.89%
SVM	48.15%
Ranbom Forest	55.56%
HDClassif	61.11%

Machine Learning using Big Data

Clustering ML of delicious coffee

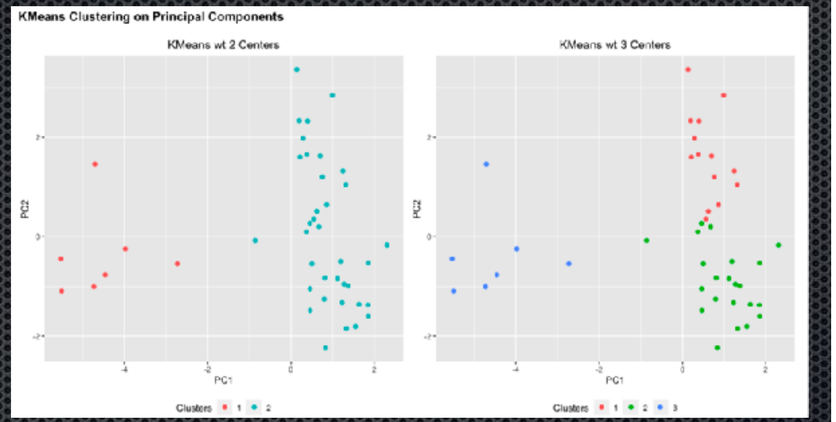
1) Hierarchical Clustering

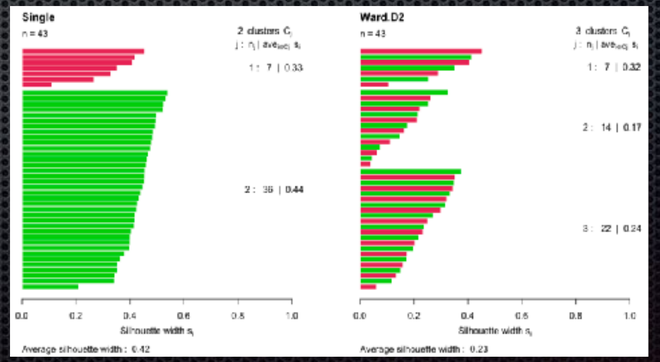


Euclidean Distance

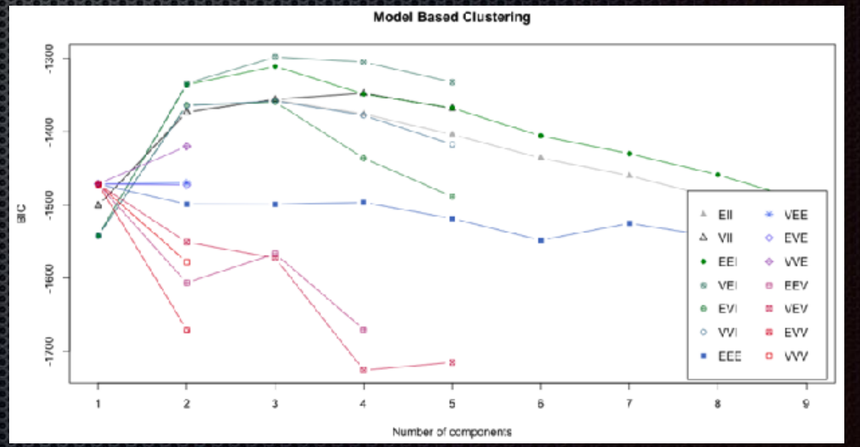
Single			Ward.D			Ward.D2		
	1	2		1	2		1	2
1	36	0	1	23	0	1	28	0
2	0	7	2	13	0	2	8	0
			3	0	7	3	0	7

2) K-Means Clustering





3) Model Based Clustering



Machine Learning using Big Data

Which one is the best, though?

Adjusted Rand Index (ARI)

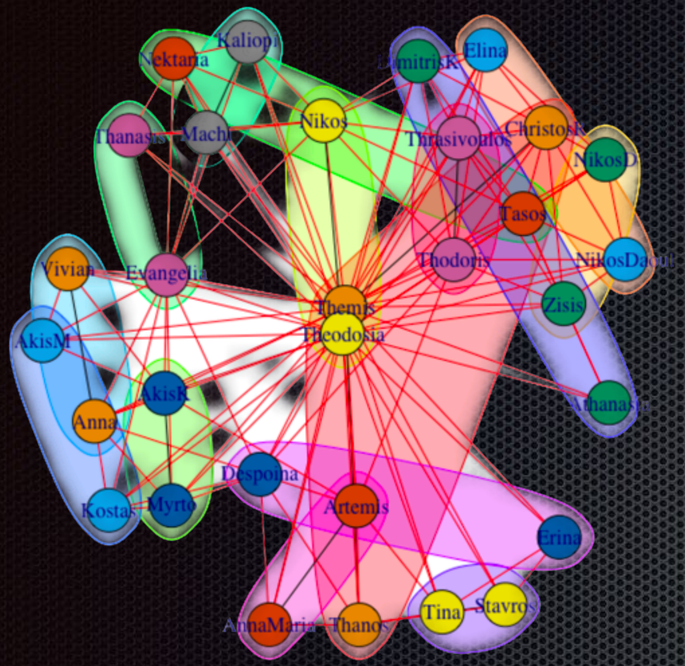
$$ARI = \frac{(\text{agreement}) - (\text{expected agreement due to chance})}{(\text{max agreement}) - (\text{expected agreement due to chance})}$$

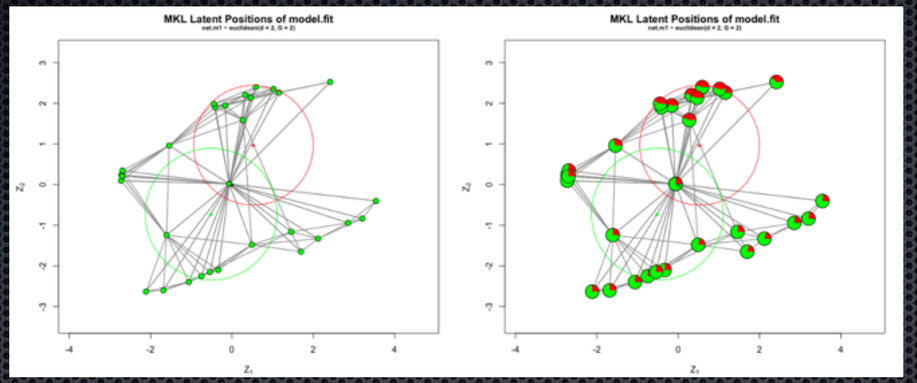
Adjusted Rand Index				
Distance	Single	Complete	Ward.D	Ward.D2
Euclidean	1	-	0.397	0.515
Manhattan	1	0.383	0.383	0.383
Minkowski	1	-	0.397	0.515

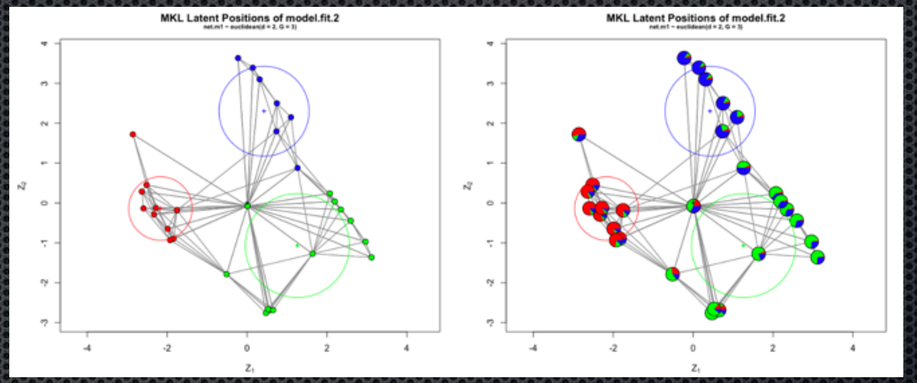
K-Means ARI	
Single	1
Ward.D2	0.383

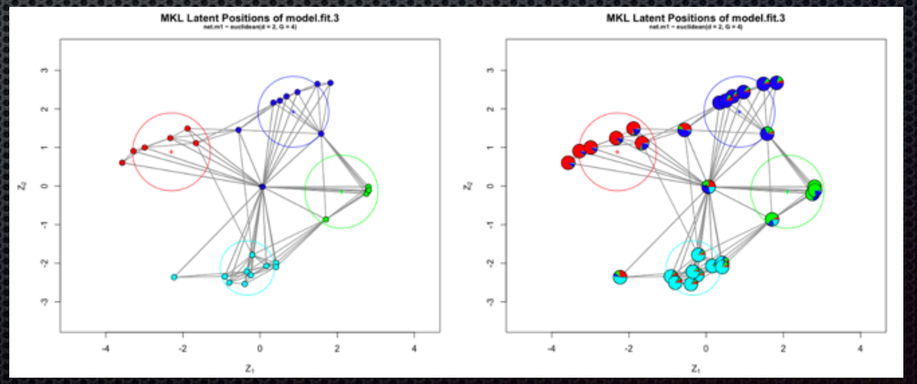
Machine Learning using Big Data

With friends, at last!









Models Comparison			
Model	BIC	L-BIC	LSpace/ClBIC
G = 2	588.06	329.3	258.76
G = 3	578.34	330.69	247.65
G = 4	551.76	330.31	221.45
G = 5	578.3	330.06	248.23