RUTGERS
Rutgers Business School
Newark and New Brunswick

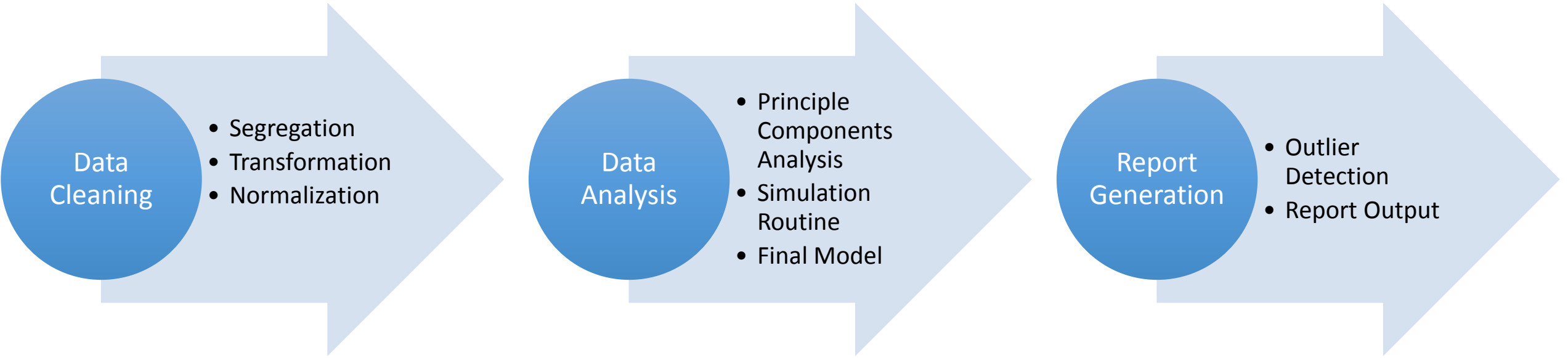
The Application of a Super App to Operational Risk

Paul Byrnes

Miklos Vasarhelyi

Yunsen Wang

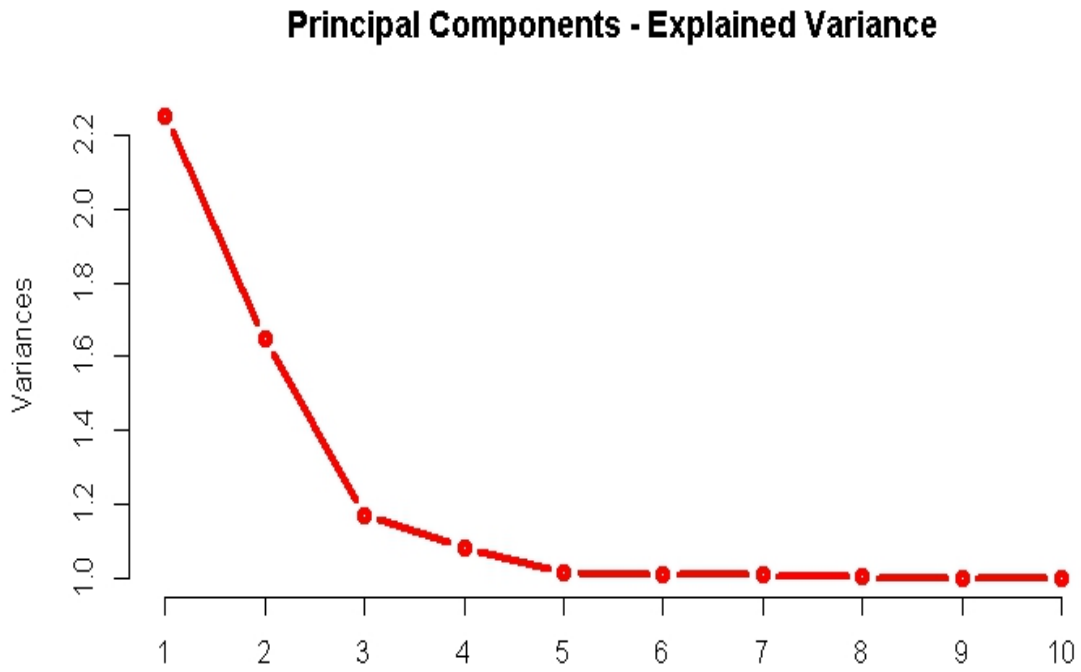
Process Outline



RUTGERS
Rutgers Business School
Newark and New Brunswick

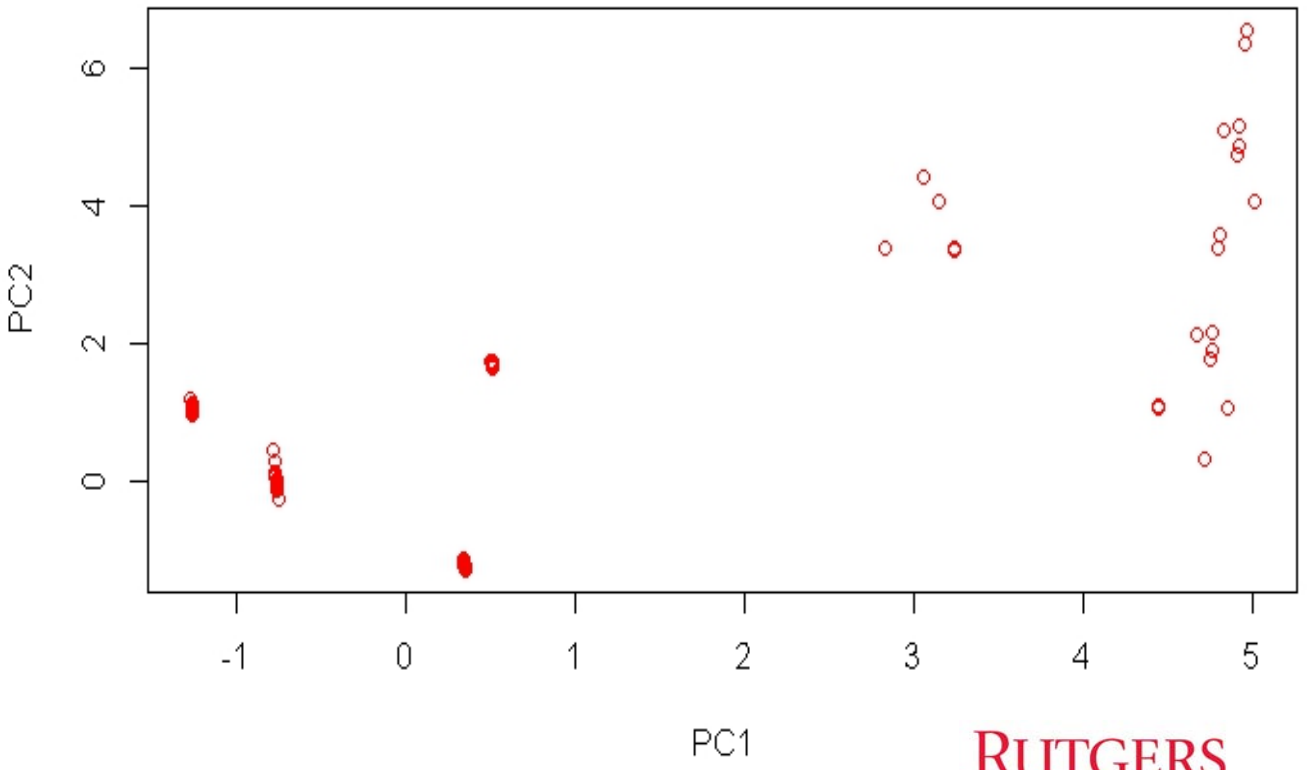
Principal Component Analysis (1/2)

- A total of 15 principal components were initially generated.
- However, based upon explained data variability threshold considerations, a subset of only 10 Principal Components are ultimately needed for analysis.



Principal Component Analysis (2/2)

- In this view, PC1 and PC2 are plotted for each object in two-dimensional space.
- From this viewpoint, records seem to be segregated into at least three major clusters.



Clustering Exploration Process

Ward Method	• Ward suggested a general agglomerative hierarchical clustering procedure where the criterion for choosing a pair of clusters to merge at each step is based on the optimal value for an objective function.
Complete Link Method	• The method is also known as farthest neighbor clustering. The result of the clustering can be visualized as a dendrogram, which shows the sequence of cluster fusion and the distance at which each fusion took place.
PAM (Partitioning Around Medoids)	• The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoids shift algorithm. More stable than k-means because it uses the median as a basis for clustering.
K-means	• k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
Expectation Maximization	• EM is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

Simulation Results – Final Model Selection

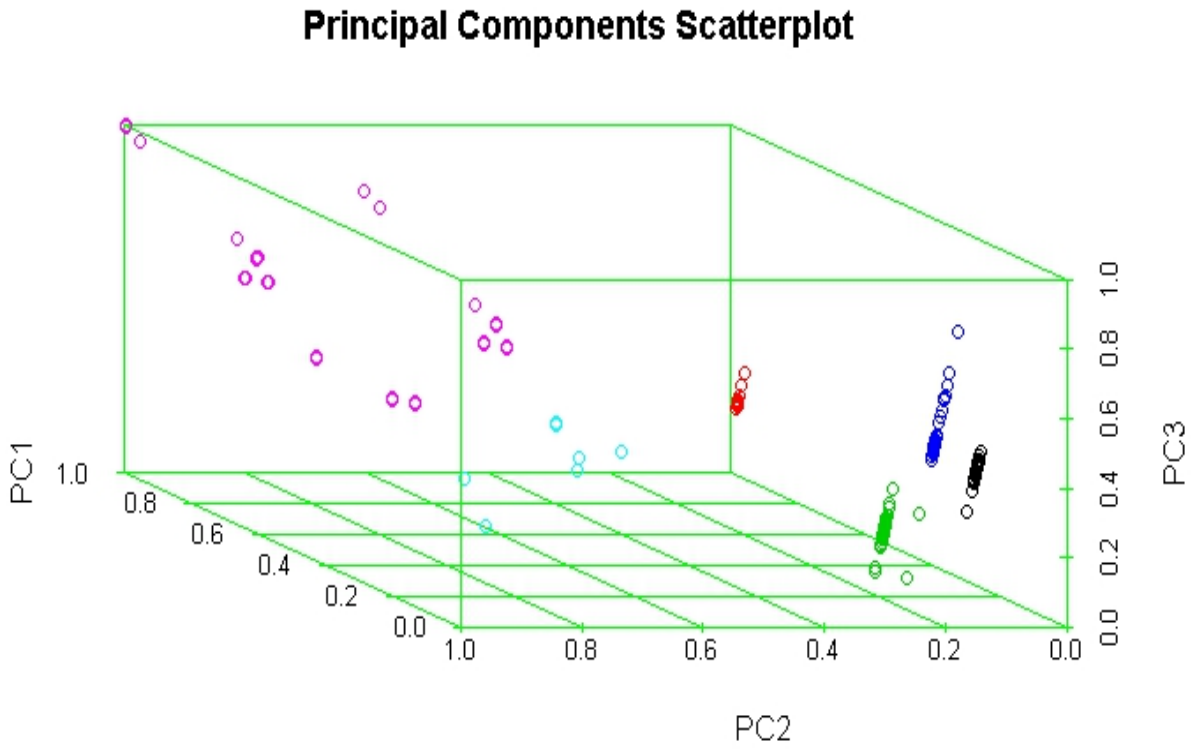
- The result below follows from a simulation routine whereby all five algorithms were compared on the data set. Furthermore, the silhouette coefficient is used as the metric for model selection. In this particular case, the preferred algorithm is Ward's method and the desired number of clusters is 6.

Best Method	Number Of Clusters	Silhouette Value
Ward's Method	6	0.962903

- The silhouette value can theoretically range from -1 to +1. It is a simultaneous measure of cohesion and separation, and, thus, is a suitable indicator of cluster quality. Incidentally, a higher silhouette value is preferred.

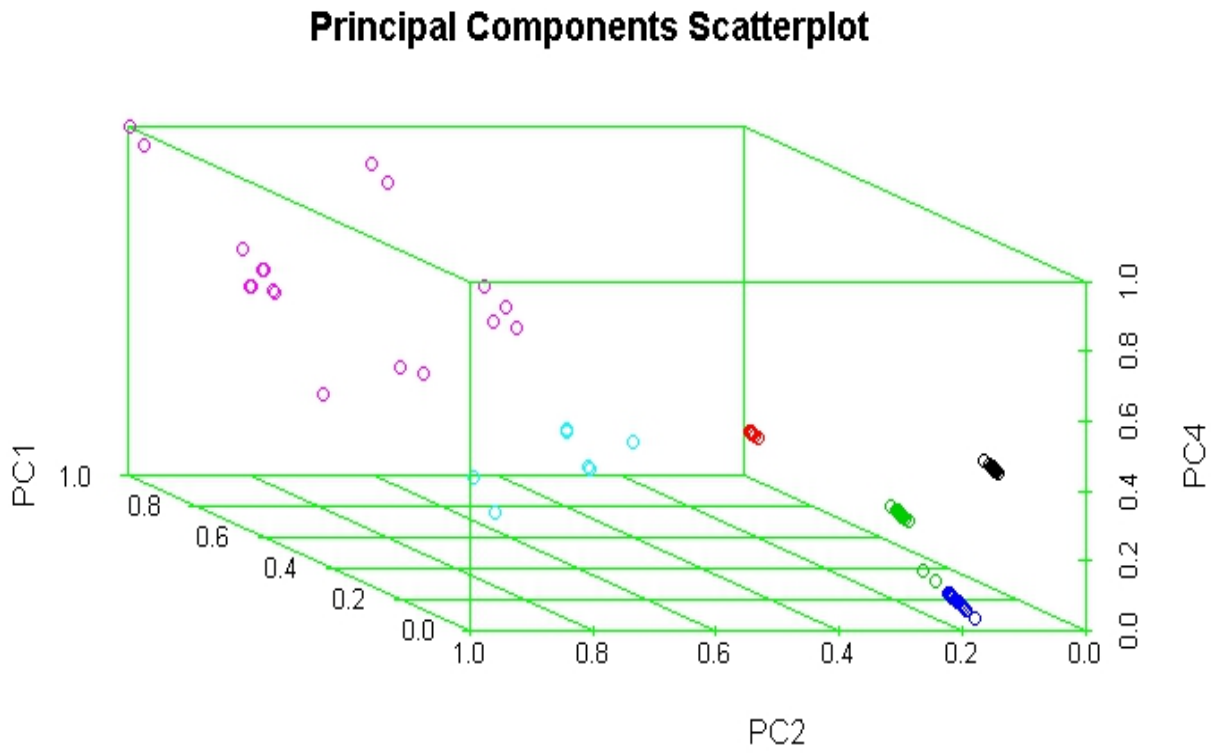
PCs with Cluster Assignment Labels (1/3)

- 3D view of normalized PC1, PC2, and PC3 of all objects.
- In this view, each cluster is depicted as a unique color and six clusters exist. From this view, one cluster appears quite large and sparse (purple objects), while the others are relatively more dense.



PCs with Cluster Assignment Labels (2/3)

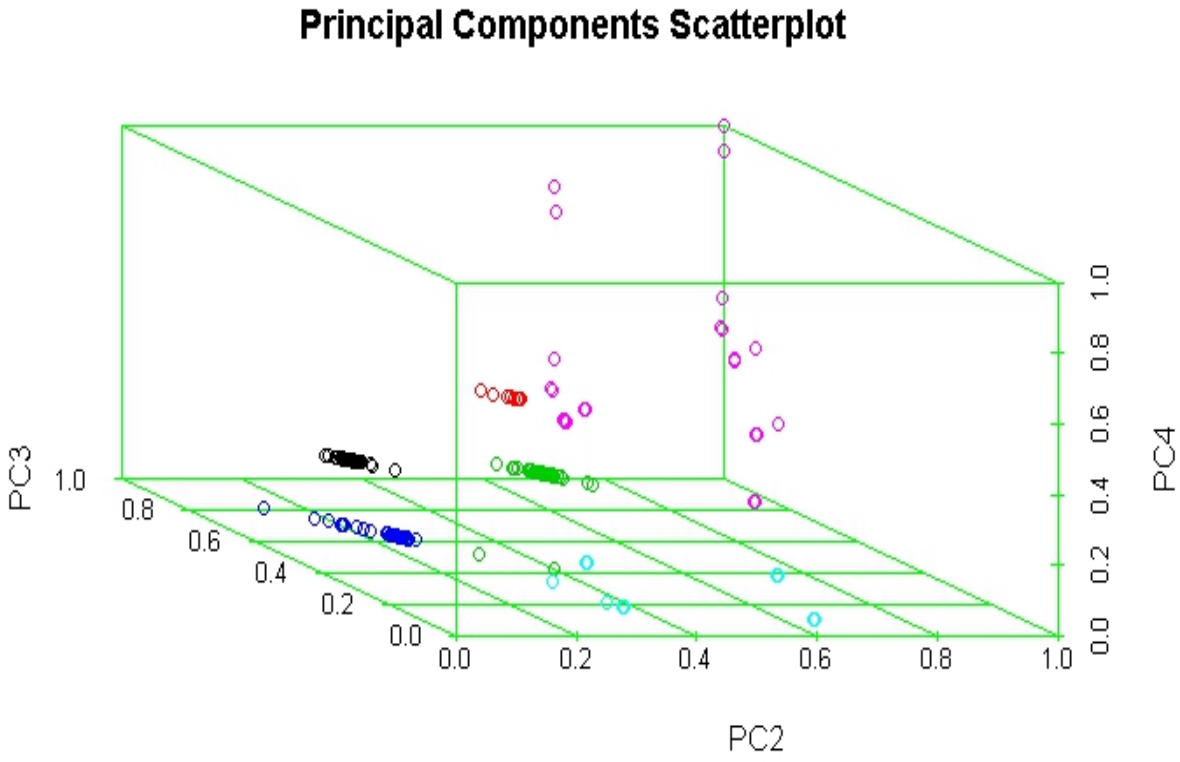
- 3D view of normalized PC1, PC2, and PC4 of all objects.
- Again, each cluster is depicted as a unique color. Observations mirror those in the previous plot.



RUTGERS
Rutgers Business School
Newark and New Brunswick

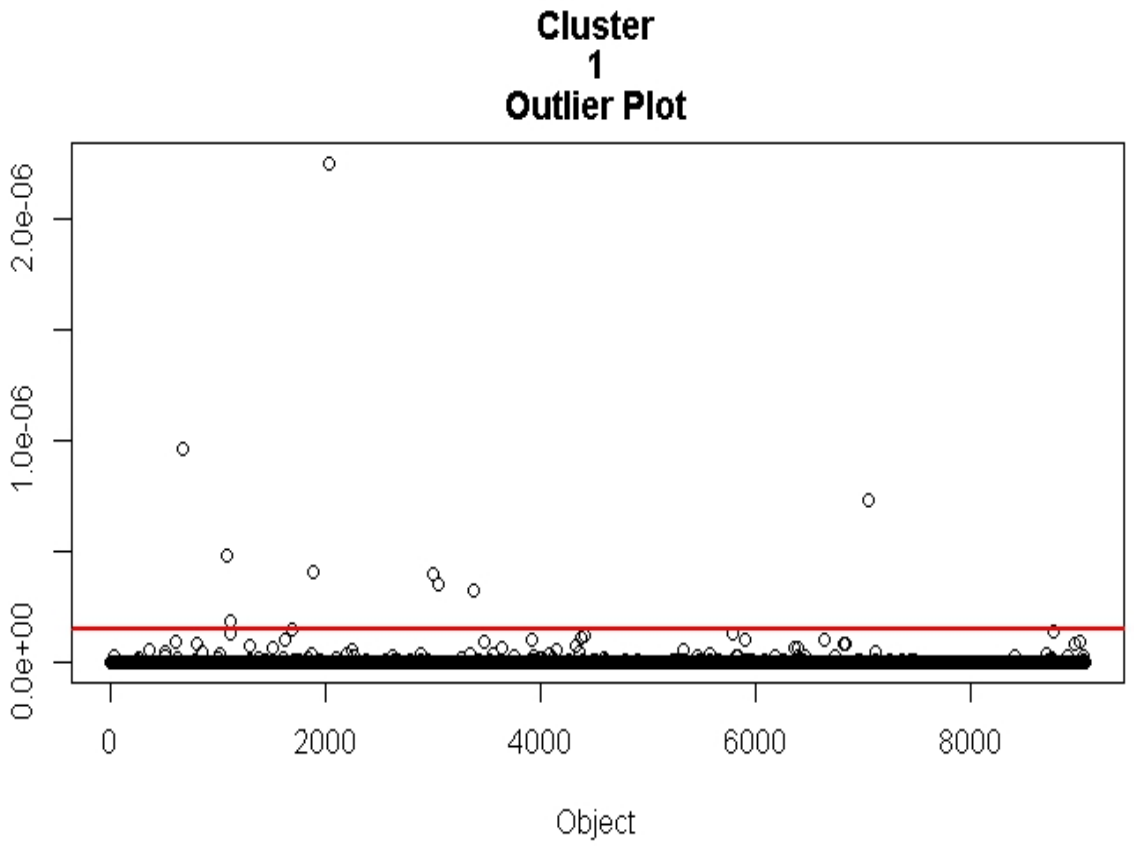
PCs with Cluster Assignment Labels (3/3)

- 3D view of normalized PC2, PC3, and PC4 of all objects.
- This provides an additional perspective about how the records are aggregated into 6 clusters. While perhaps not as clear as the previous two graphs, it again depicts one cluster as much larger than the others in terms of both region size and sparsity.



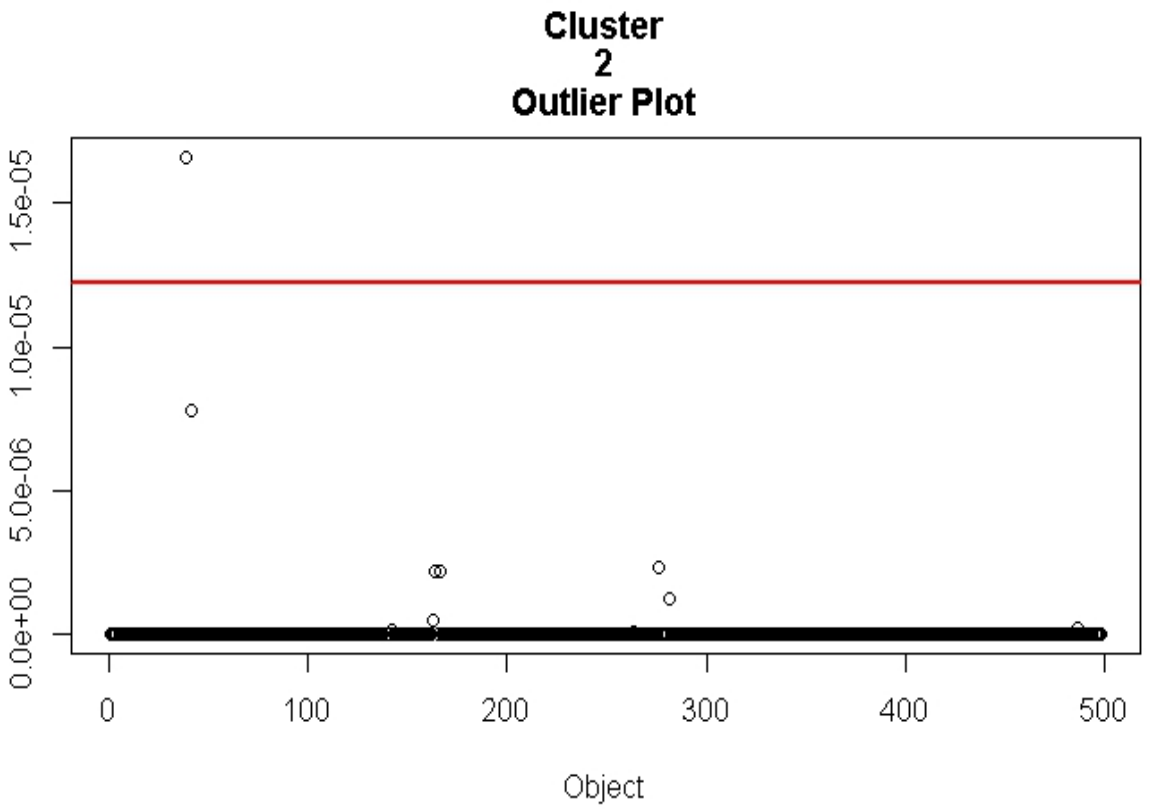
Cluster 1 Outlier

- Objects beyond the cutoff would be viewed as particularly suspicious.
- The following rows in cluster 1 data appear very suspicious:
  - 678 (ID. 11838020)
  - 1086 (ID. 11921859)
  - 1119 (ID. 11921906)
  - 1683 (ID. 11987844)
  - 1888 (ID. 11988241)
  - 2032 (ID. 12104041)
  - 2994 (ID. 12232567)
  - 3044 (ID. 12232644)
  - 3371 (ID. 12491049)
  - 7053 (ID. 41030079)



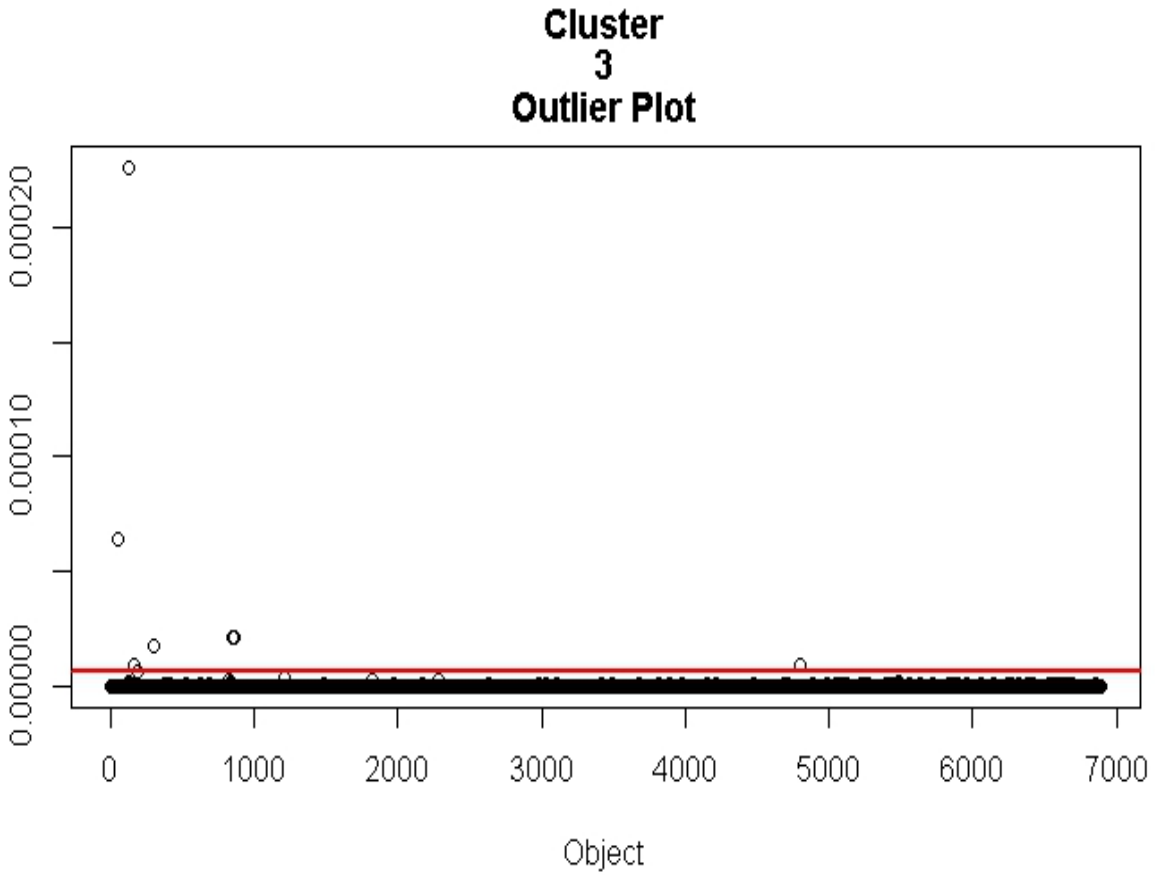
Cluster 2 Outlier

- Objects beyond the cutoff would be viewed as particularly suspicious.
- The following rows in cluster 2 data appear very suspicious:
  - 39 (ID. 7447449)



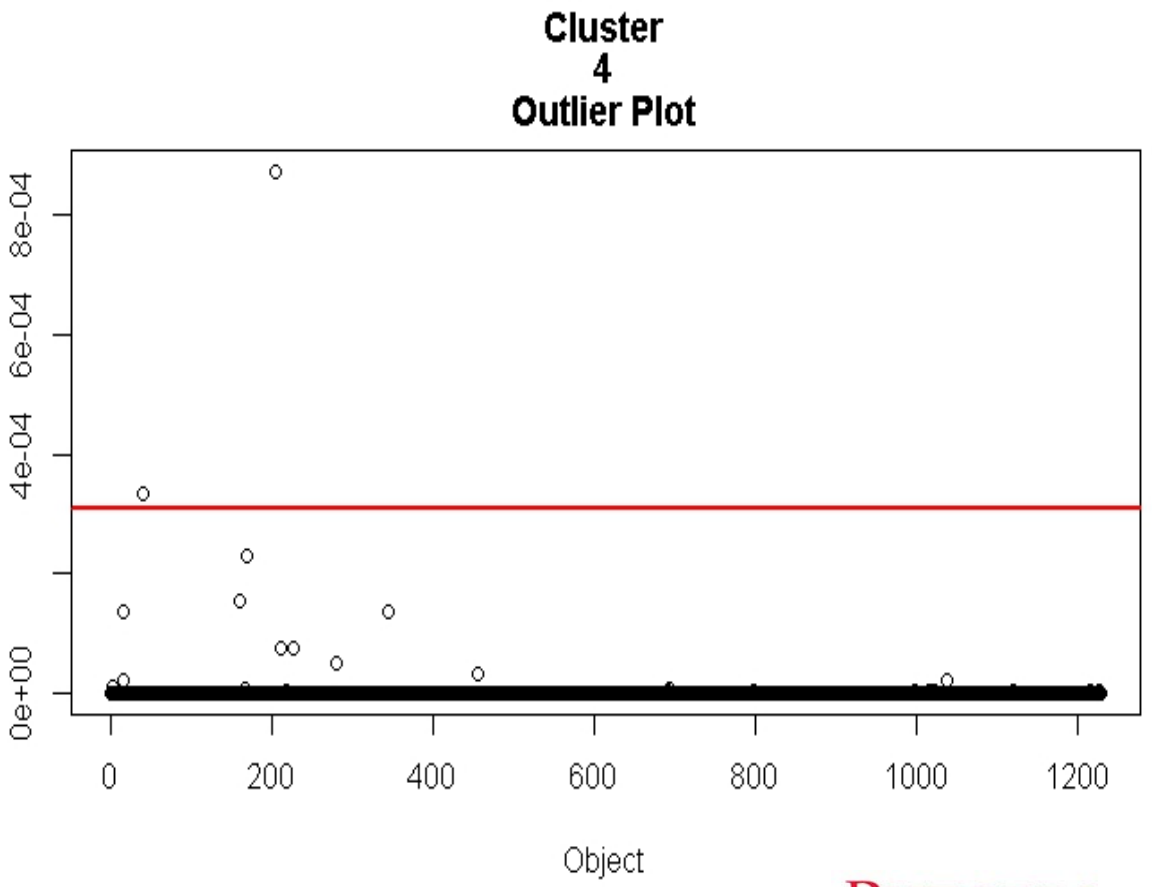
Cluster 3 Outlier

- Objects beyond the cutoff would be viewed as particularly suspicious.
- The following rows in cluster 3 dat a appear very suspicious:
  - 55 (ID. 7446822)
  - 120 (ID. 11987940)
  - 159 (ID. 12104704)
  - 302 (ID. 13767107)
  - 850 (ID. 23290275)
  - 858 (ID. 23290291)
  - 4804 (ID. 41030759)



Cluster 4 Outlier

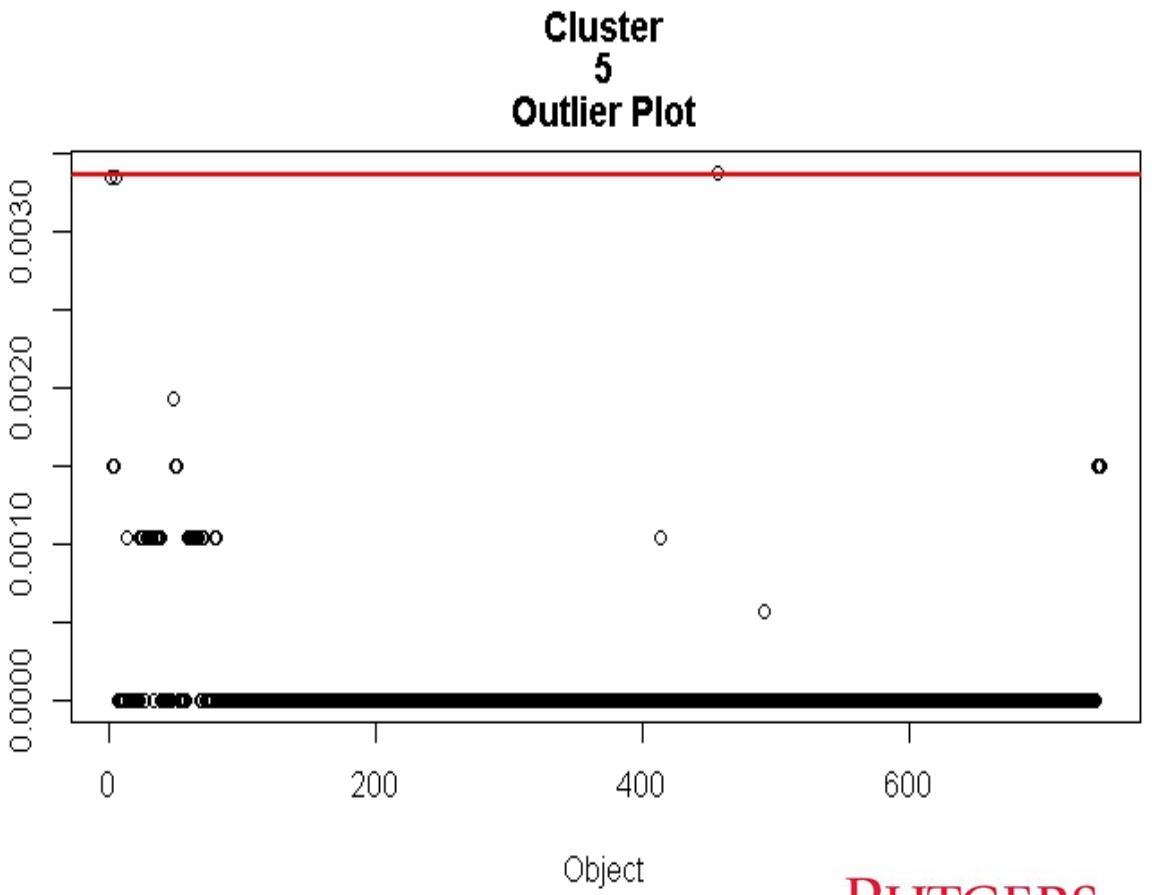
- Objects beyond the cutoff would be viewed as particularly suspicious.
- The following rows in cluster 4 data appear very suspicious:
  - 41 (ID. 7446820)
  - 204 (ID. 11987934)



RUTGERS
Rutgers Business School
Newark and New Brunswick

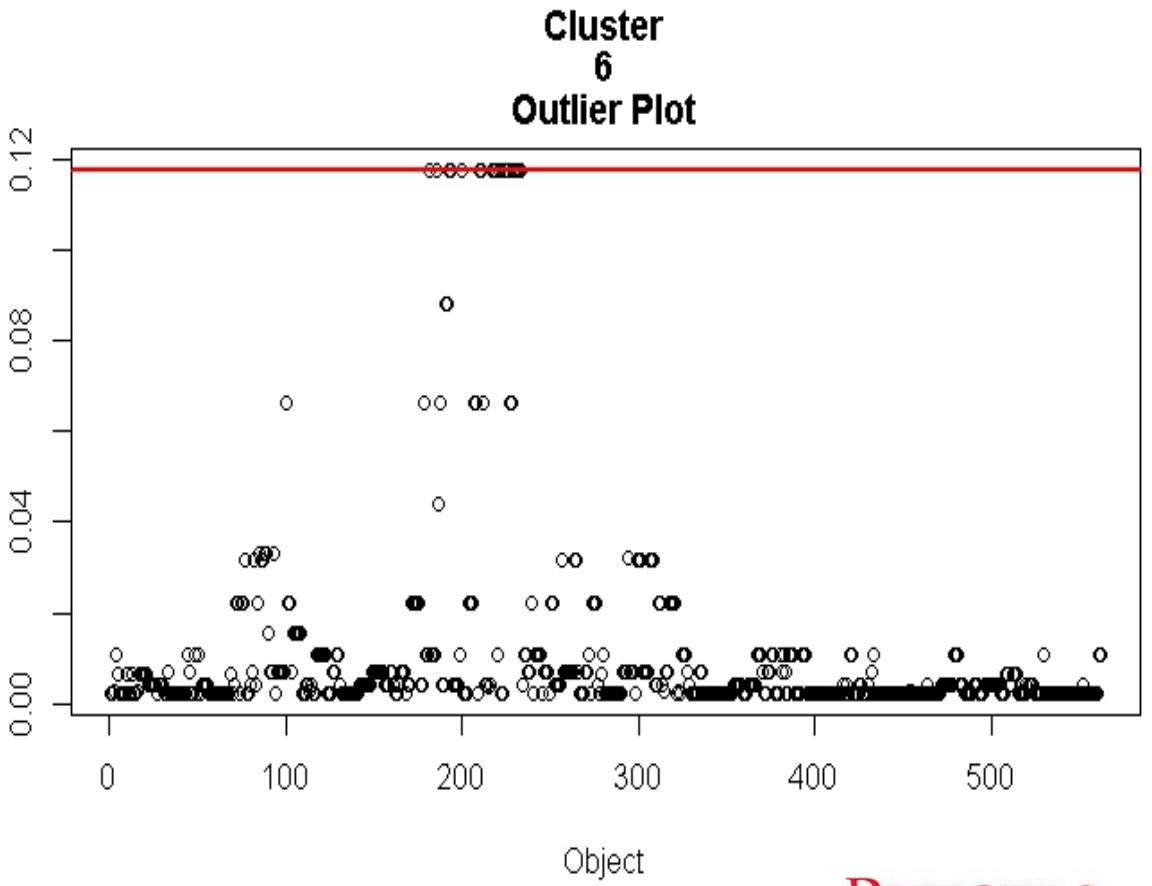
Cluster 5 Outlier

- Objects beyond the cutoff would be viewed as particularly suspicious.
- The following rows in cluster 5 data appear very suspicious:
  - 457 (ID. 43341470)



Cluster 6 Outlier

- Objects beyond the cutoff would be viewed as particularly suspicious.
- No rows in cluster 6 data appear highly suspicious. However, several exist on the threshold and might warrant further study. These records are:
  - 217 (ID. 41029713)
  - 218 (ID. 41029714)
  - 221 (ID. 41029717)
  - 182 (ID. 41029648)
  - 186 (ID. 41029652)
  - 193 (ID. 41029663)
  - 210 (ID. 41029698)
  - 211 (ID. 41029699)
  - 224 (ID. 41029724)
  - 225 (ID. 41029725)
  - 226 (ID. 41029726)
  - 229 (ID. 41029730)
  - 230 (ID. 41029731)
  - 233 (ID. 41029734)



RUTGERS
Rutgers Business School
Newark and New Brunswick

Report Generation

- At the end of outlier detection, an output file for each cluster containing the record identifier, original variables, normalized variables, principal components, normalized principal components, cluster assignment labels, and Mahalanobis distance information is exported in CSV format file to facilitate further analyses and investigations.

Thank you!

Yunsen Wang