Feature Extraction and Labelling Large Data Sets Using Deep Learning

Dr. Daqing Chen

Division of Computer Science and Informatics

School of Engineering

London South Bank University

ST$_i$VE2017 - RESEARCHER LINK: Smart Technology for fighting Zika-Virus Epidemics
10 - 13 September 2017
University of Pernambuco,● Recife-PE ● Brazil ●
http://www.graphicslink.co.uk/STfVE2017/

Introduction

- There are many high-volume, high-dimensional historical data sets that need to be analysed, e.g., The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute of the USA: Cancer patients' records for over 40 years with more than 10 million instances.
- Such data sets are usually featured with categorical data type and therefore result in a data set of high-dimensionality and high-sparsity
- Challenges: How to group and label such data sets effectively?

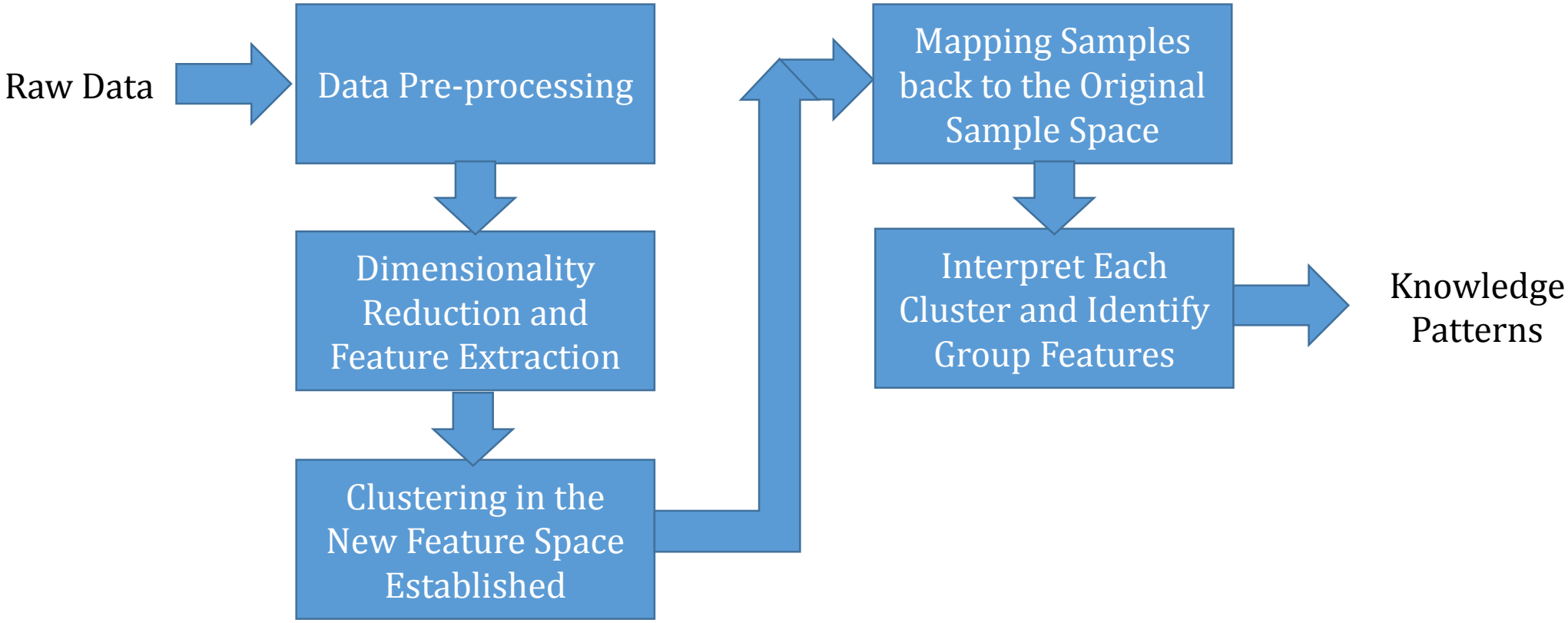
Methods for Grouping/Labelling Data

- The k-means clustering algorithm and its variants
  - Based on a certain similarity measure, e. g. the Euclidian distance between samples.
  - Difficult to apply to a data set of high-dimensionality and high-sparsity.
  - Sensitive to noisy.
- Dimensionality reduction using, for example, PCA, SVD, and then grouping data in the feature space
  - Difficult to apply to a data set of high-dimensionality.
  - The dimensionality of the feature space may still be very high, e. g., if the variables are less correlated or even independent on each other

Using Deep Learning for Effective Feature Extraction and Dimensionality Reduction

- Restrict Boltzmann Machines (RBMs)
  - A paradigm of deep learning networks (artificial neural networks).
  - Can effectively transform data from a original sample space to a new feature space of a very low-dimensionality.
  - Can be used as a group in sequence, and the outputs of one RBM are used as the inputs to the next RBM. This enables the feature of the data can be extracted gradually and the number of the dimensions in a new feature space can be controlled.

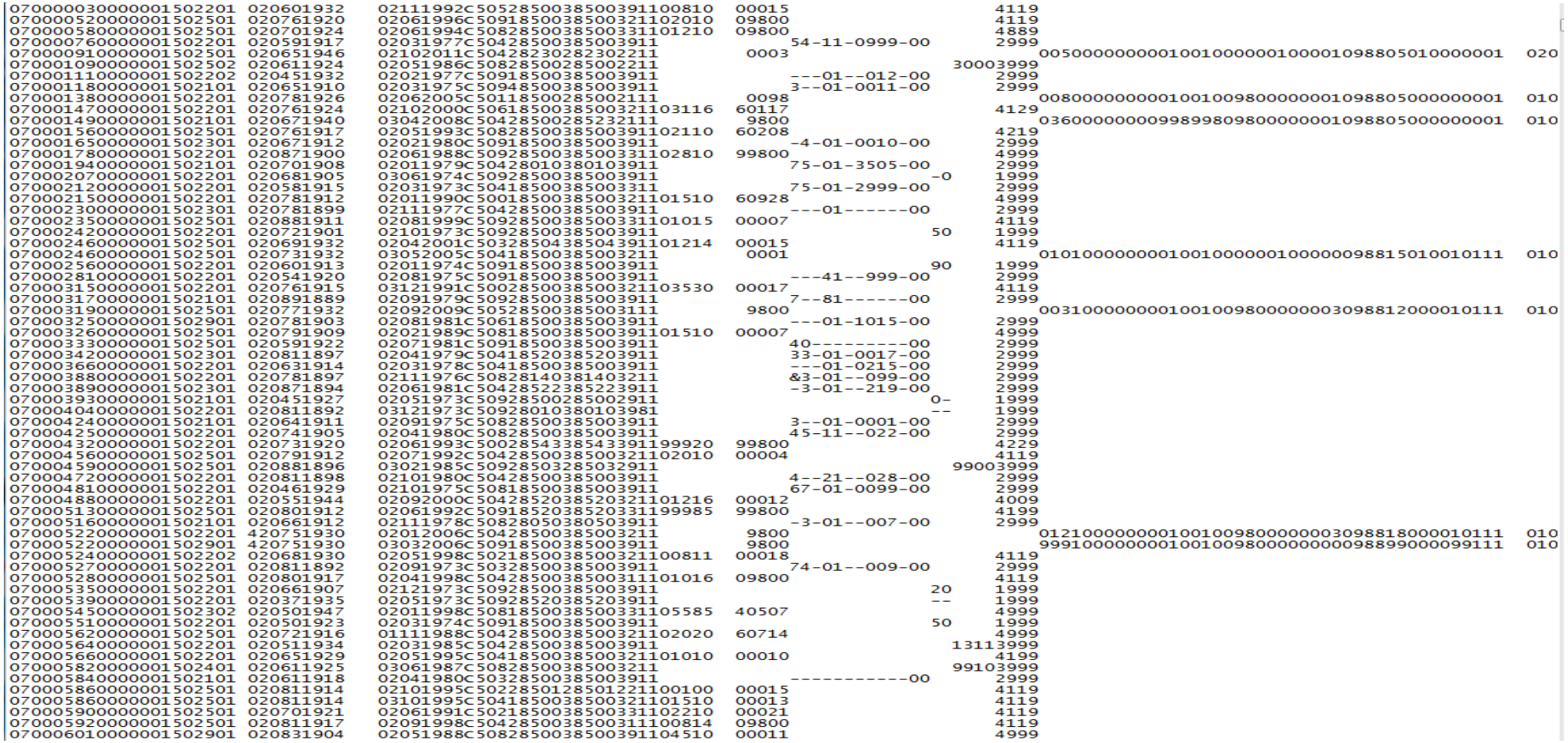
The Proposed Approach



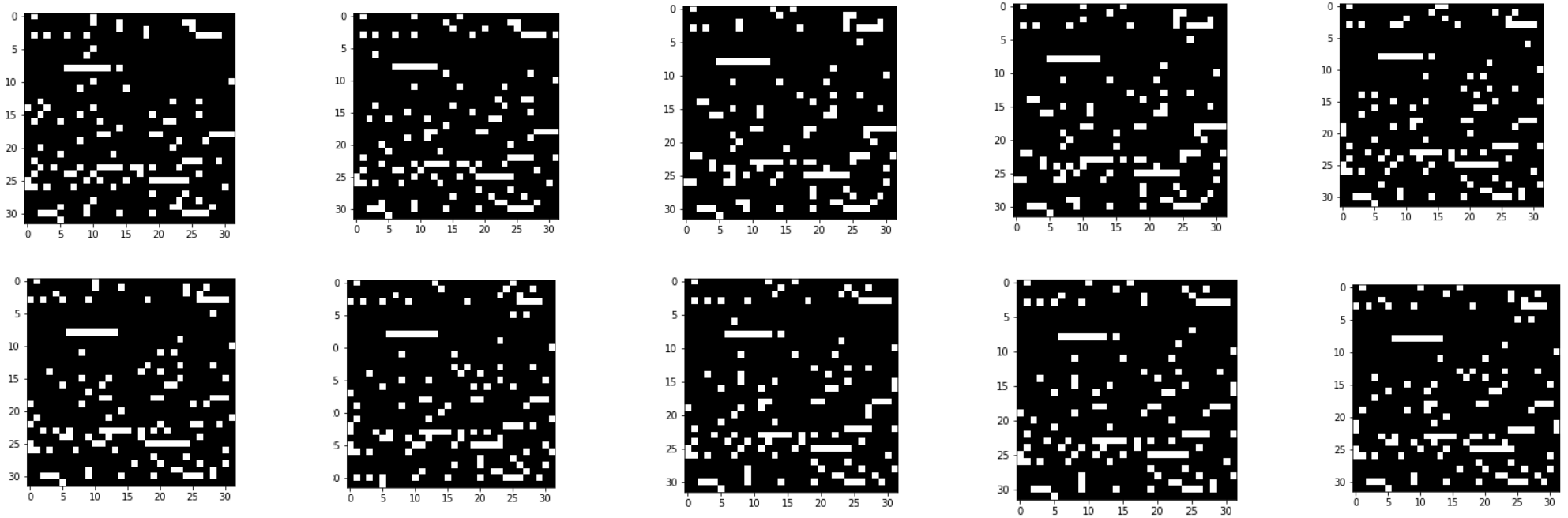
An Example

- The raw data: the SEER breast cancer data: 260K instances with 130 variables (most of them are categorical type)
- Data pre-processing
  - Convert the original data using the one-hot method (i.e., orthogonal coding, or dummy encoding): each of the distinct values of a given categorical variable forms a new column. This has resulted in1006 dimensions in the target data set.
  - Dealing with missing values.
  - Normalising data.
  - Presenting the data as an image.
- RBMs: Two used, 33×33 (original) → 25×25 (new space 1) →9×9 (new space 2)
- The k-means clustering applied to the outputs of the last RBM, 6 clusters created.
- Map samples back to the original space and Interpret each cluster in terms of similarity and diversity.

The Original Raw Data

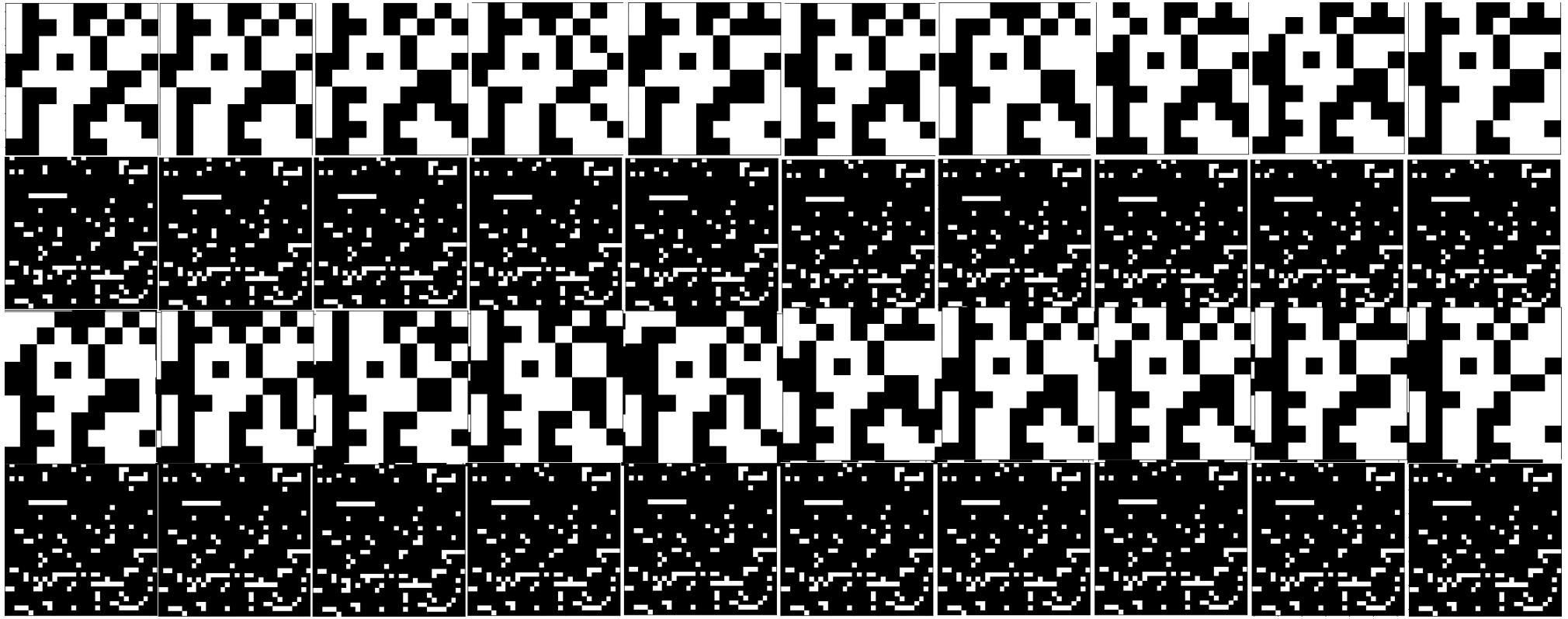


The Target Data Presented as a Set of Images

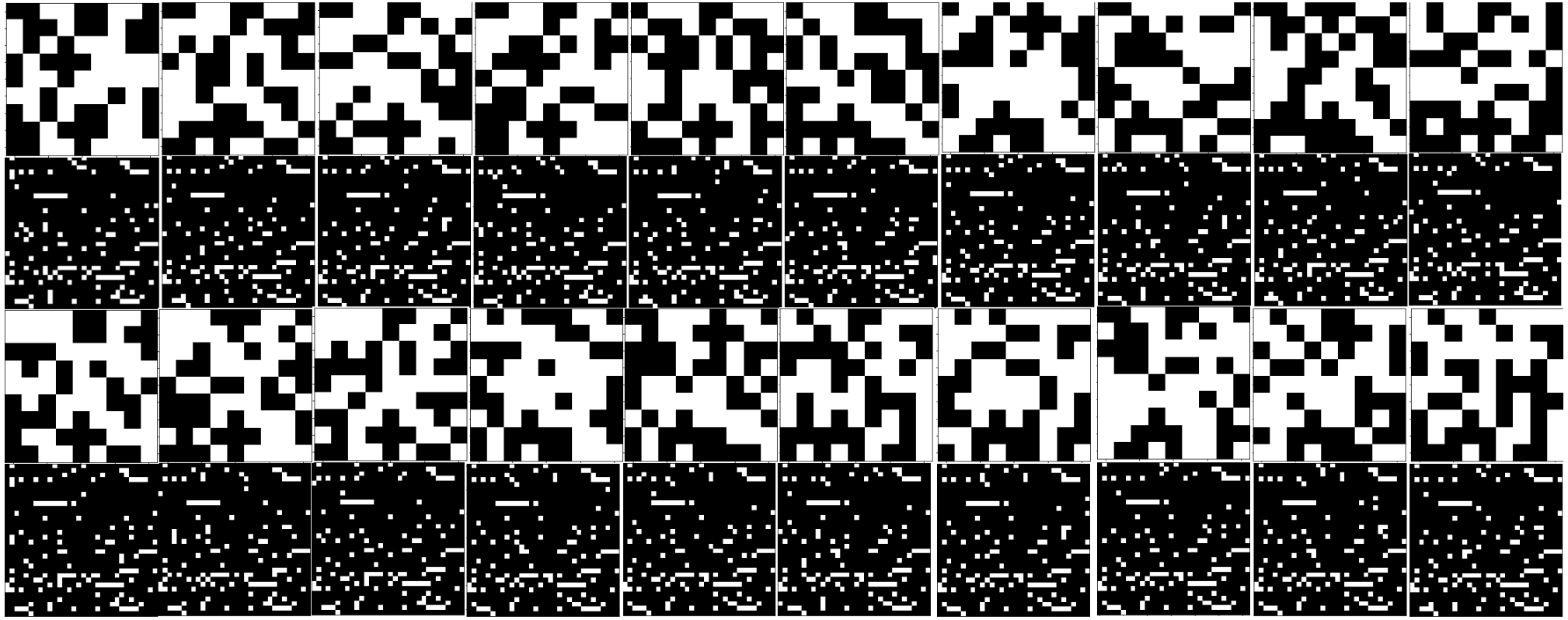


Each image represents a patient

Results



Results



Results

Clusters	Cluster 0	Cluster 1	Cluster 3	Cluster 5	Cluster 4	Cluster 2
Survival Rate	25.00 %	92.00 %	89.00 %	78.00 %	74.00 %	72.00 %
Grade						
Grade 1	7.00 %	23.00 %	14.00 %	32.00 %	3.00 %	10.00 %
Grade 2	27.00 %	45.00 %	31.00 %	59.00 %	8.00 %	41.00 %
Grade 3	31.00 %	24.00 %	27.00 %	3.00 %	80.00 %	43.00 %
Grade 4	3.00 %	0.00 %	2.00 %	0.00 %	3.00 %	1.00 %
Cell not determined	32.00 %	8.00 %	26.00 %	6.00 %	6.00 %	5.00 %
	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
Treatment						
Surgery performed	38.00 %	95.00 %	95.00 %	99.00 %	99.00 %	99.00 %
Surgery not recommanded	34.00 %	2.00 %	3.00 %	1.00 %	1.00 %	1.00 %
Contraindicated	2.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Died before	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Unknown reason for no surgery	15.00 %	1.00 %	2.00 %	0.00 %	0.00 %	0.00 %
Refused	4.00 %	1.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Recommanded	1.00 %	1.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Unknown if surgery performed	6.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %

Conclusion

- RBMs can be effectively integrated into a big data analytics process.
- The similarity features in the original space can be effectively extracted by using RBMs, brought forward, and presented in the feature space established.
- This approach can be applied to other data sets including government data.

References

- G. E. Hinton and R. R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, SCIENCE, VOL. 313, 28 JULY 2006.
- D. Chen and E. Zhao, Deep Learning Big Social Data: An Analysis on the Diversity and Similarity of London Wards, LSBU Research Report, July 2017.
- D. Chen, S. Mallet, and P. Camenen, Deep Learning Casual Attributions for Breast Cancer, LSBU Research Report, August 2017.

Key questions for debate in this session

- What features does a RBM extract?
- How to interpret the feature space established by a RBM?
- Design an ideal imaginary description?