

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

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?

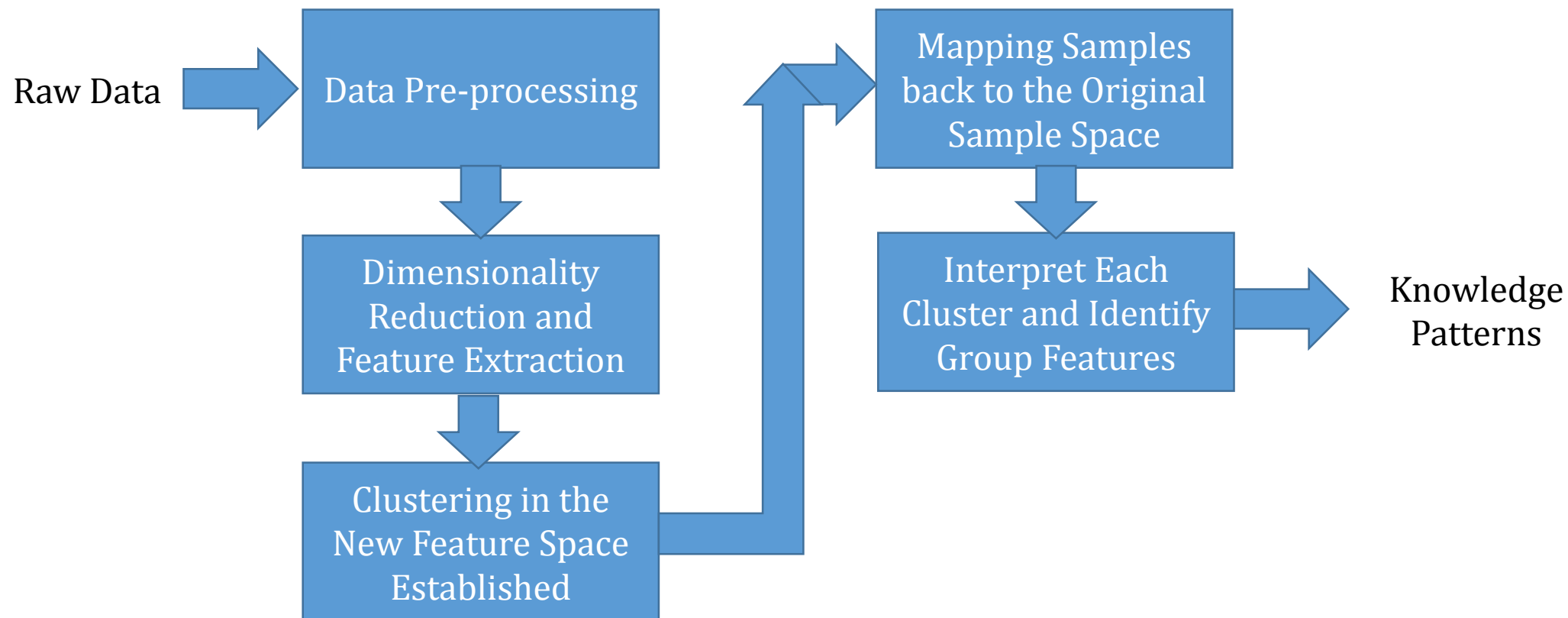
Methods for Grouping/Labeling Data

- 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 in 1006 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.

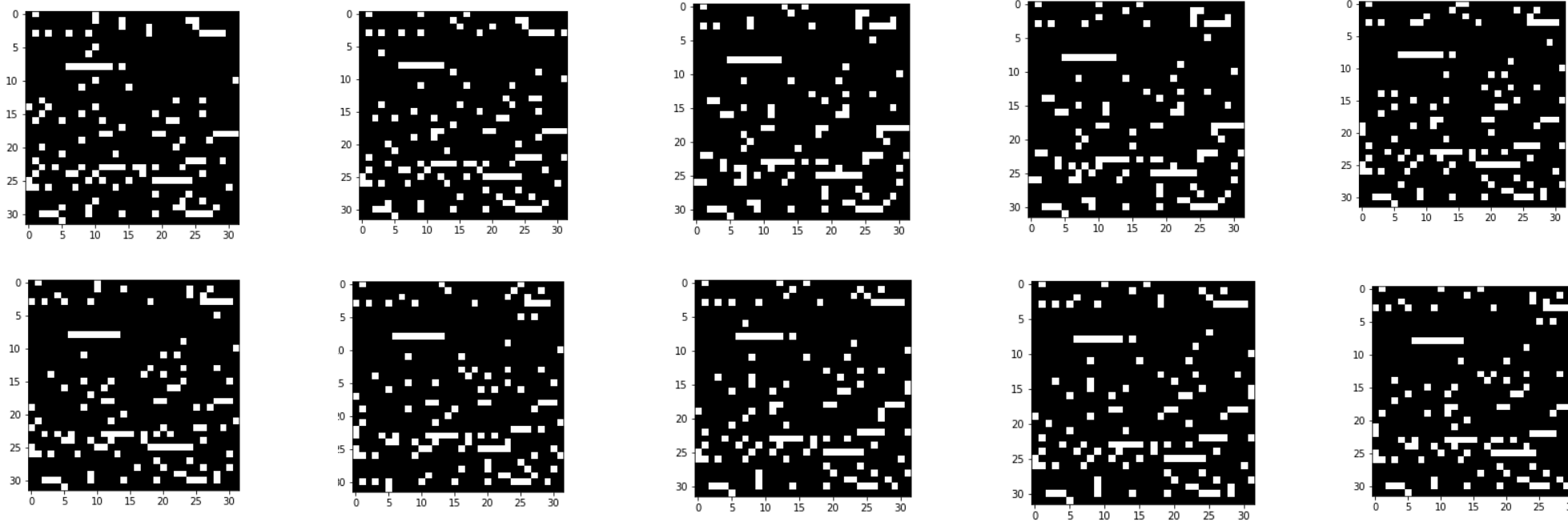
The Original Raw Data

```

0700000300000001 502201 020601932 02111992C 5052850038500391100810 00015 4119
0700005200000001 502501 020761920 02061996C 5091850038500321102010 09800 4119
0700005800000001 502501 020701924 02061994C 5082850038500331101210 09800 4888
0700007600000001 502201 020591917 02031977C 50428500385003911 0003 2999
0700009100000001 502501 020651946 02102011C 50428230282302211 0003 30003999
0700010900000001 502502 020611924 02051986C 50828500285002211 0003 0050000000001001000000100001098805010000001 020
0700011100000001 502202 020451932 02021977C 50918500385003911 0003 2999
0700011800000001 502101 020651910 02031975C 50948500385003911 0003 2999
0700013800000001 502201 020781926 02062005C 50118500285002111 0098 0080000000001001009800000001098805000000001 010
0700014700000001 502201 020761924 02102000C 5061850038500321103116 60117 4129
0700014900000001 502101 020671940 03042008C 50428500285232111 9800 0360000000099899809800000001098805000000001 010
0700015600000001 502501 020761917 02051993C 5082850038500391102110 60208 4219
0700016500000001 502301 020671912 02021980C 50918500385003911 0003 2999
0700017800000001 502201 020871900 02061988C 5092850038500331102810 99800 4999
0700019400000001 502101 020701908 02011979C 50428010380103911 0003 2999
0700020700000001 502201 020681905 03061974C 50928500385003911 0003 1999
0700021200000001 502201 020681915 02031973C 504118500385003311 0003 2999
0700021500000001 502201 020781912 02011990C 5001850038500321101510 60928 4999
0700023000000001 502301 020781899 02111977C 50428500385003911 0003 2999
0700023500000001 502501 020881911 02081999C 5092850038500331101015 00007 4119
0700024200000001 502201 020721901 02101973C 50928500385003911 0003 50
0700024600000001 502501 020691932 02042001C 5032850438504391101214 00015 1999
0700024600000001 502501 020731932 03052005C 50418500385003211 0001 4119
0700025600000001 502201 020601913 02011974C 50918500385003911 0003 90
0700028100000001 502201 020541920 02081975C 50918500385003911 0003 1999
0700031500000001 502201 020761915 03121991C 5002850038500321103530 00017 2999
0700031700000001 502101 020691889 02091979C 50928500385003911 0003 4119
0700031900000001 502501 020771932 02092009C 50528500385003311 9800 2999
0700032500000001 502901 020781903 02081981C 50618500385003911 9800 0031000000001001009800000003098812000010111 010
0700032600000001 502501 020791909 02021989C 5081850038500391101510 00007 2999
0700033300000001 502501 020591922 02071981C 50918500385003911 0003 4999
0700034200000001 502301 020811897 02041979C 50418520385203911 40-----00 2999
0700036600000001 502201 020631914 02031978C 50418500385003911 33-01-0017-00 2999
0700038800000001 502201 020781897 02111976C 50828140381403211 ---01-0215-00 2999
0700038900000001 502301 020871894 02061981C 50428522385223911 &3-01--099-00 2999
0700039300000001 502101 020451927 02051973C 50928500285002911 -3-01--219-00 1999
0700040400000001 502201 020811892 03121973C 50928010380103981 0-----00 1999
0700042400000001 502101 020641911 02081975C 50828500385003911 3--01-0001-00 1999
0700043200000001 502201 020741905 02041980C 50828500385003911 45-11--022-00 2999
0700043200000001 502201 020731920 02061993C 5002854338543391199920 99800 4229
0700045600000001 502501 020791912 02071992C 5042850038500321102010 00004 4119
0700045900000001 502501 020881896 03021985C 50928503285032911 99003999
0700047200000001 502201 020811898 02101980C 50428500385003911 4--21--028-00 2999
0700048100000001 502201 020461929 02101975C 50818500385003911 67-01-0099-00 2999
0700048800000001 502201 020551944 02092000C 5042852038520321101216 00012 4009
0700051300000001 502501 020801912 02061992C 5091852038520331199985 99800 4199
0700051600000001 502101 020661912 02111978C 50828500380503911 -3-01--007-00 2999
0700052200000001 502201 420751930 02012006C 50428500385003211 9800 0121000000001001009800000003098818000010111 010
0700052400000001 502201 020751930 03032008C 50918500385003911 9800 9991000000001001009800000000098899000099111 010
0700052200000001 502202 020681930 02051998C 5021150038500321100811 00018 4119
0700052700000001 502201 020811892 02091973C 50328500385003911 00018 2999
0700052800000001 502501 020801917 02041998C 5042850038500311101016 09800 4119
0700053500000001 502201 020661907 02121973C 50928500385003911 02121973C 50928500385003911 20 1999
0700053900000001 502201 020371935 02051973C 50928520385203911 -- 1999
0700054500000001 502302 020501947 02011998C 5081850038500331105585 40507 4999
0700055100000001 502201 020501923 02031974C 50918500385003911 50 1999
0700056200000001 502501 020721916 01111988C 5042850038500321102020 60714 4999
0700056400000001 502201 020511934 02031985C 50428500385003911 13113999
0700056600000001 502201 020651929 02051995C 5041850038500321101010 00010 4199
0700058200000001 502401 020611925 0209058200000001 502401 020611925 03061987C 50828500385003211 99103999
0700058400000001 502101 020611918 02041980C 50328500385003911 -----00 2999
0700058600000001 502501 020811914 02101995C 5022850128501221100100 00015 4119
0700058600000001 502501 020811914 03101995C 5041850038500321101510 00013 4119
0700059000000001 502501 020701921 02061991C 5021850038500331102210 00021 4119
0700059200000001 502501 020811917 02091998C 5042850038500311100814 09800 4119
0700060100000001 502901 020831904 02051988C 5082850038500391104510 00011 4999

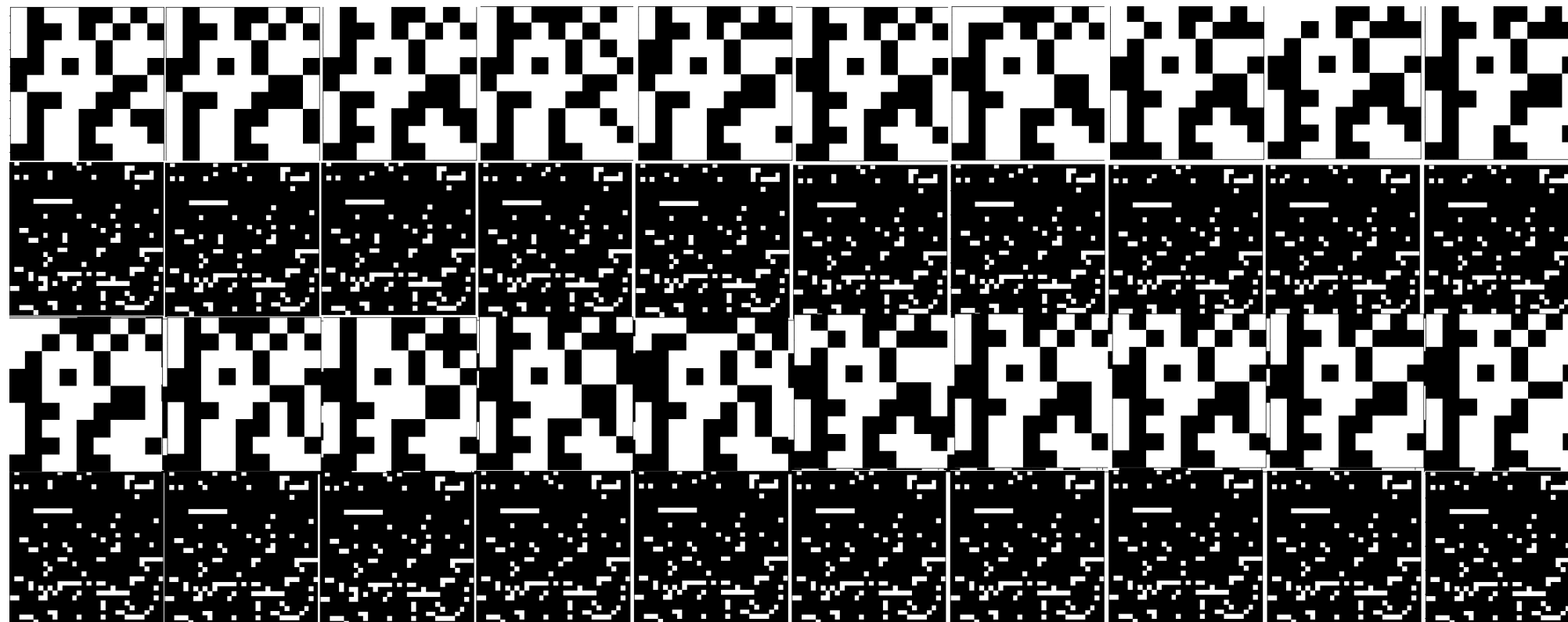
```

The Target Data Presented as a Set of Images

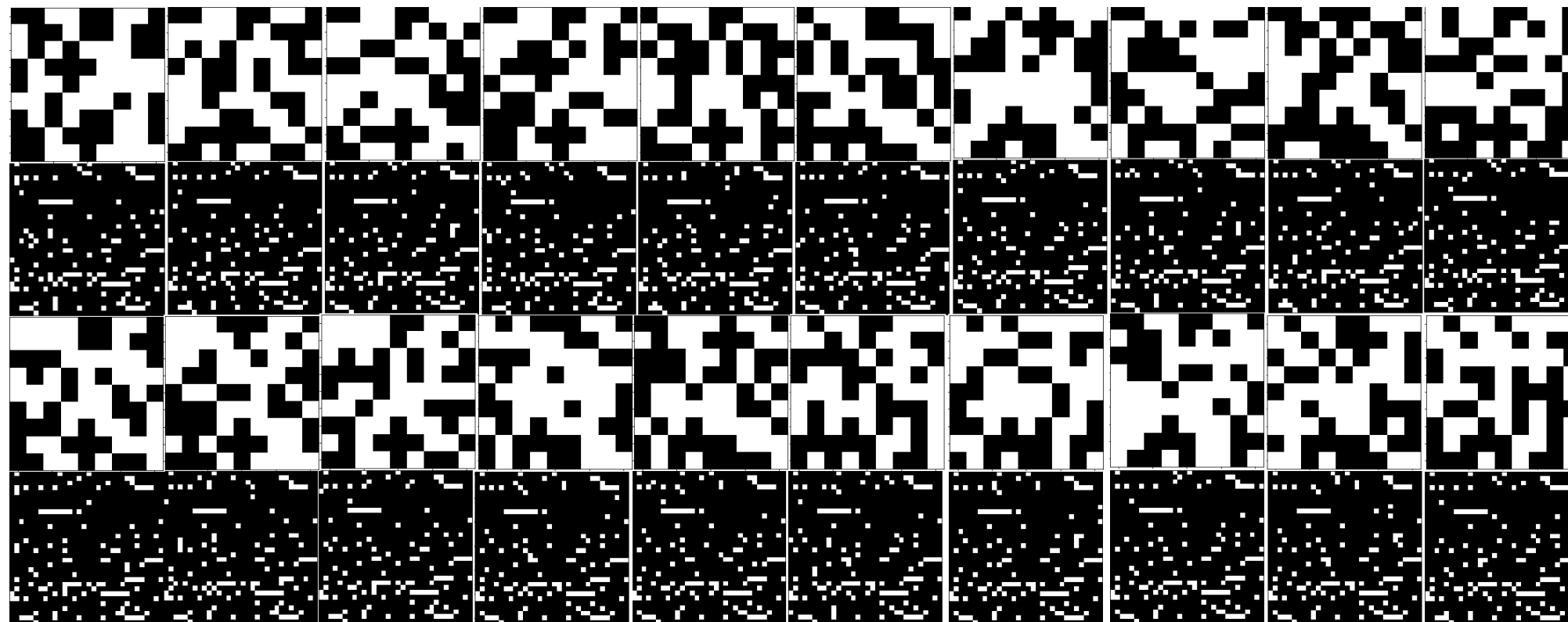


Each image represents a patient

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 recommended	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 %
Recommended	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?