Visualizing Hidden Structures in Datasets using Deep Learning

Dmytro Perekrestenko

Supervisor: Xavier Bresson

Professor: Pierre Vandergheynst

June 3, 2015

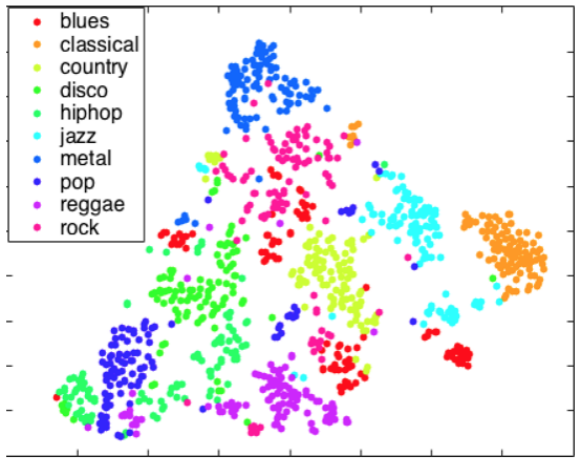
Project Goal

Develop a tool to find hidden structures in large-scale and high-dimensional datasets.

Complex high-dimensional dataset

2D Visualisation

Deep Learning
+
t-SNE



Find hidden structures

Handcrafted Audio Features [Sturm, 2013]

- **Short-term features:**
  Octave-based spectral contrast (OSC), Mel-frequency cepstral coefficients (MFCCs), spectral centroid, rolloff, flux, zero-crossings.

- **Long-term features:**
  Octave-based modulation spectral contrast (OMSC), "low-energy", modulation spectral flatness measure (MSFM) and modulation spectral crest measure (MSCM).

$\Rightarrow$ The dimension of feature representation of 30-sec song is **1150**.

Temporal Echonest Audio Features [Schindler, 2014]

- **Segments Timbre** - 12 dimensional vector with unbounded values centered around 0 representing a high level abstraction of the spectral surface.
- **Segments Pitches** - normalized 12 dimensional vector ranging from 0 to 1 corresponding to the 12 pitch classes C, C#, to B.
- **Segments Loudness Max** represents the peak loudness value within each segment.
- **Segments Loudness Max Time** describes the offset within the segment of the point of maximum loudness.
- **Segments Start** provide start time information of each segment/onset.

$\Rightarrow$ **Temporal Echonest Features (TEN)**: All statistical moments of Segments Pitches, Segments Timbre, Segments Loudness Max, Segments Loudness Max Time and lengths of segments calculated from Segments Start are calculated.

The dimension of feature representation of 30-sec song is **232**.

Unsupervised Feature Learning with Sparse Coding [Henaff et al., 2011]

Find basis (dictionary) $D = [d_1, d_2, \ldots, d_m] \in \mathbb{R}^{n \times m}$ and coefficient matrix $Z \in \mathbb{R}^{m \times N}$, such that obtained projection is ℓ_2 close to the original data matrix $X \in \mathbb{R}^{n \times N}$ and Z is sparse:

$$\min_{Z,D} \frac{1}{2}||X - DZ||_2^2 + \lambda||Z||_1, \quad \text{s.t. } ||d_j||_2 \leq 1, \ \forall j \in [1, \ldots, m].$$

We use $m = 225$ element dictionary, that means that the dimension of feature representation of 30-sec song is **225**.

Data Visualisation: t-SNE [van der Maaten, 2008]

$\Rightarrow$ **t-distributed stochastic neighbor embedding (t-SNE)** is a nonlinear dimensionality reduction technique for embedding high-dimensional data into a low-dimensional space (e.g. 2D).

The t-SNE algorithms comprises two main stages:

1. Given a set of N data points $x_1, \ldots, x_N \in \mathbb{R}^n$ ($n \gg 1$), t-SNE computes probabilities p_{ij} that are proportional to the similatrity of data points x_i and x_j:

$$p_{i|j} = e^{-\frac{||x_i - x_j||^2}{2\sigma_i^2}}, \; p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}.$$

2. t-SNE defines a similar probability distribution over the points $y_1, \ldots, y_N \in \mathbb{R}^2$ in the low-dimensional map: $q_{ij} = \frac{1}{1+||y_i - y_j||^2}$ and minimizes the Kullback–Leibler divergence between the two distributions:

$$\min_{\{y_i\}_{i=1}^N} KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

MNIST Dataset

The MNIST [LeCun et al., 1998] dataset consists of 60,000 images of handwritten digits ('0'-'9'). Image size is 28x28, dimensionality 784.

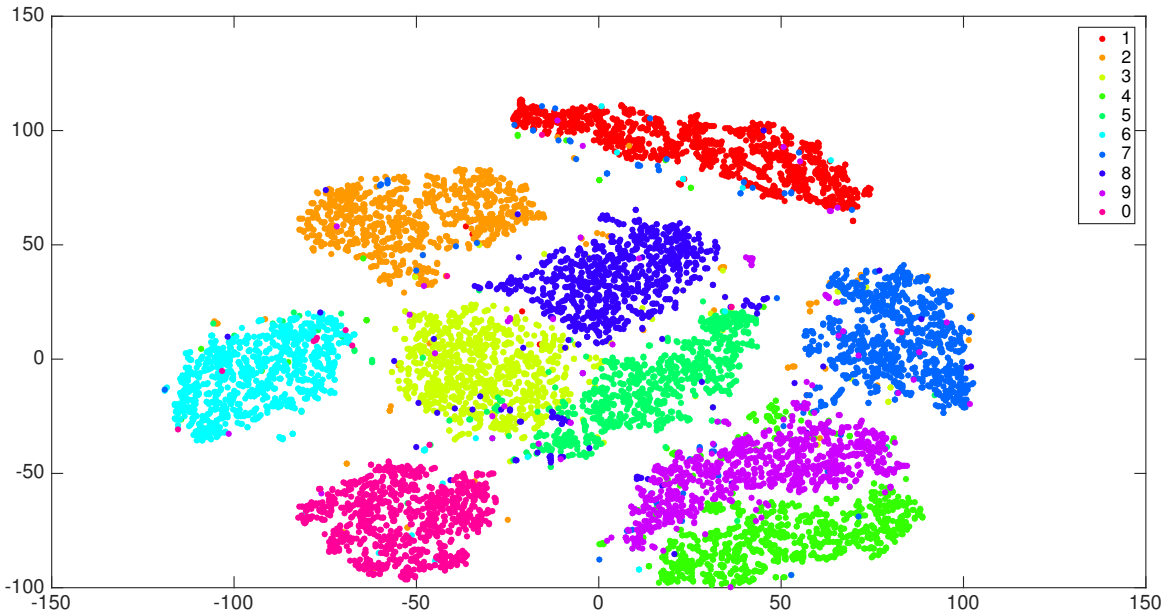


Figure: Original MNIST dataset (t-SNE projection)

MNIST Dataset

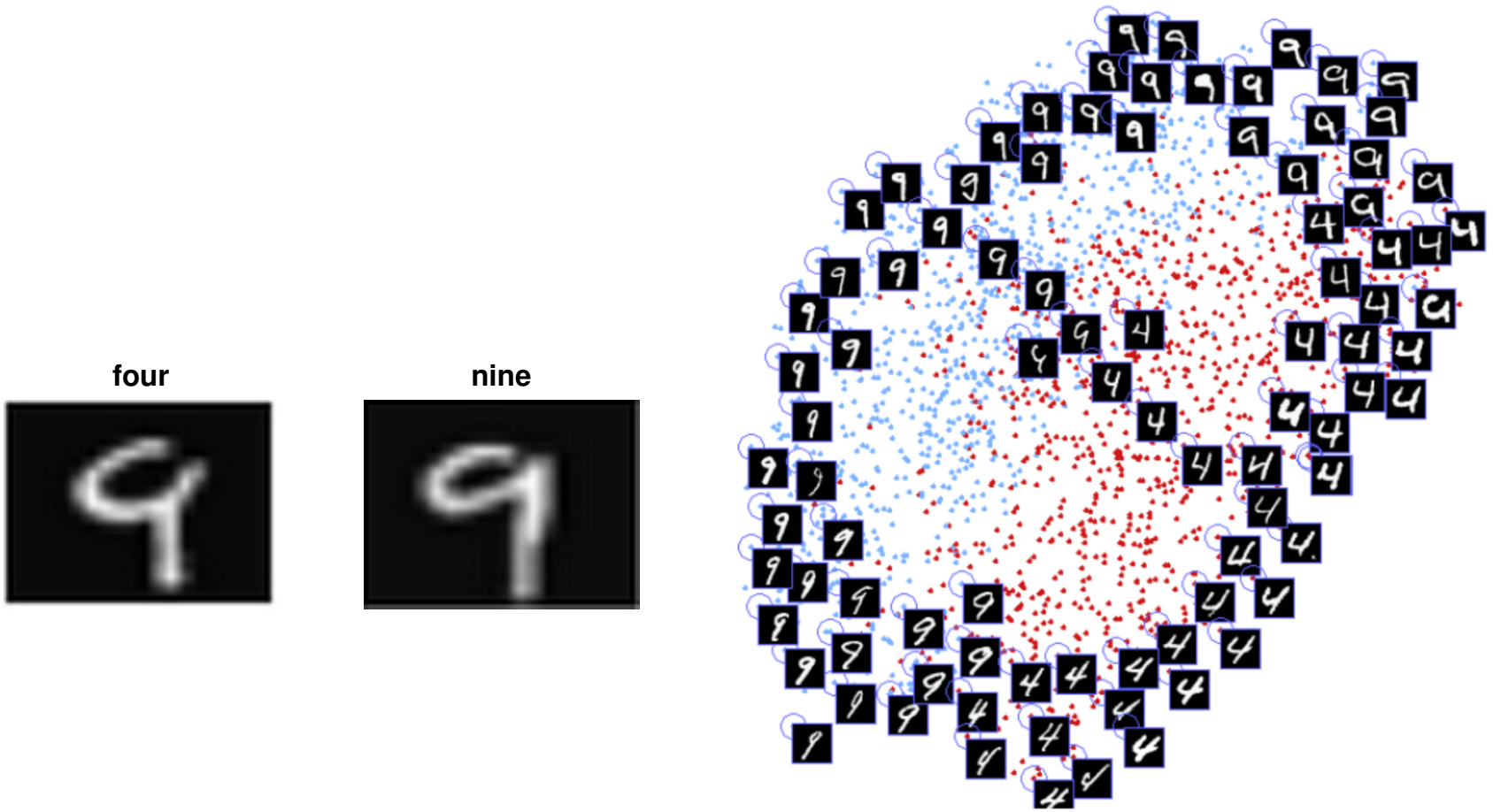


Figure: Nines and fours in MNIST dataset

GTZAN ([Sturm '13] Handcrafted Features)

The GTZAN dataset [Tzanetakis, 2002] consists of 1,000 audio tracks of 30 second length. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format.

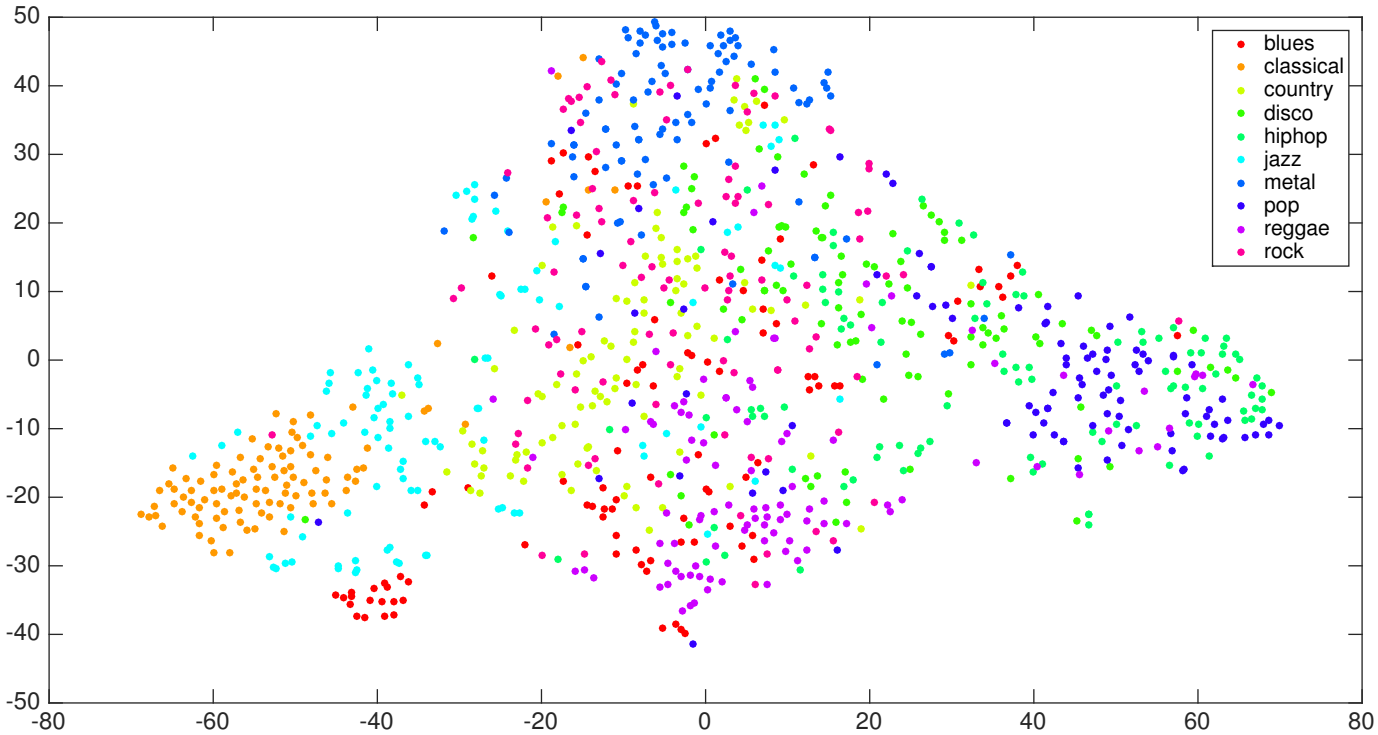


Figure: Original GTZAN dataset using [Sturm '13] features visualized by t-SNE

GTZAN (Learned Features)

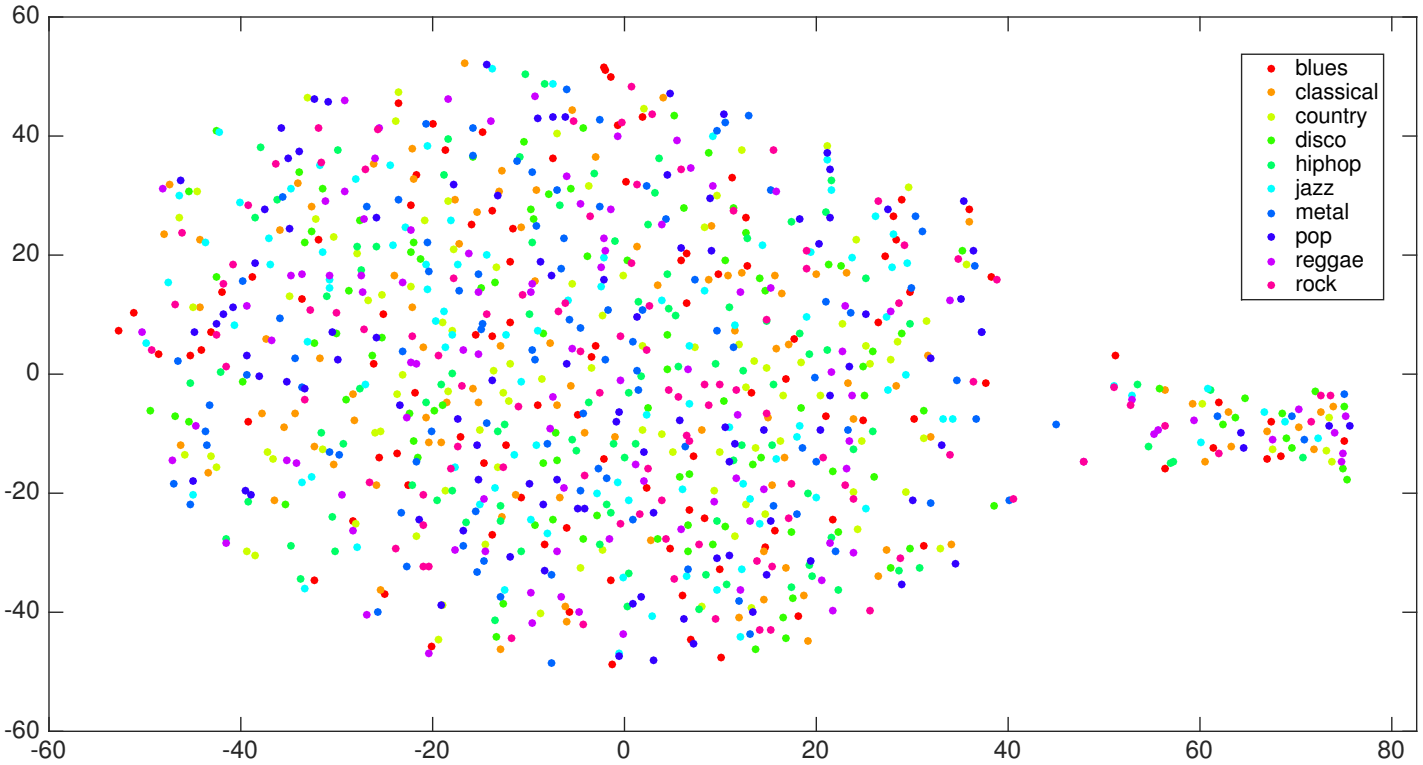


Figure: Original GTZAN dataset using learned features visualized by t-SNE

MJF Dataset

The Montreux Jazz Festival [http://www.montreuxjazz.com/] dataset consists of 9,912 samples. We extracted a subset of 3,726 samples to have 9 balanced classes (genres) with 414 songs in each. We considered for each song its temporal echonest features.

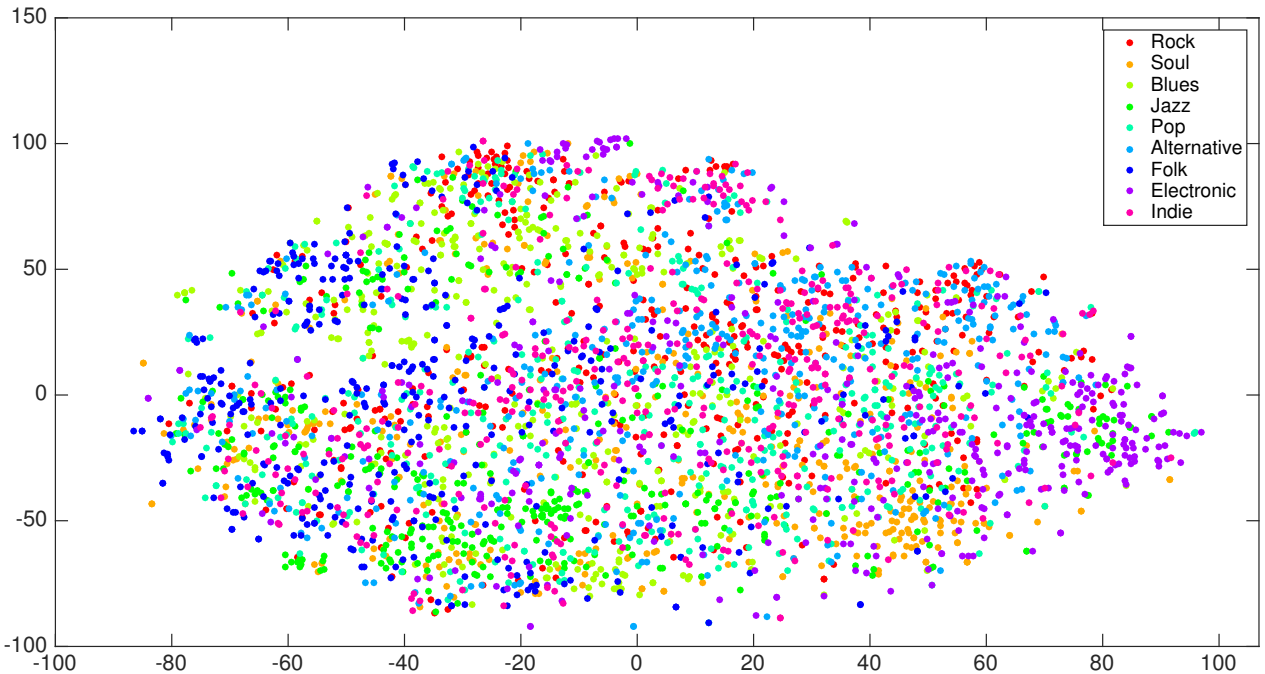
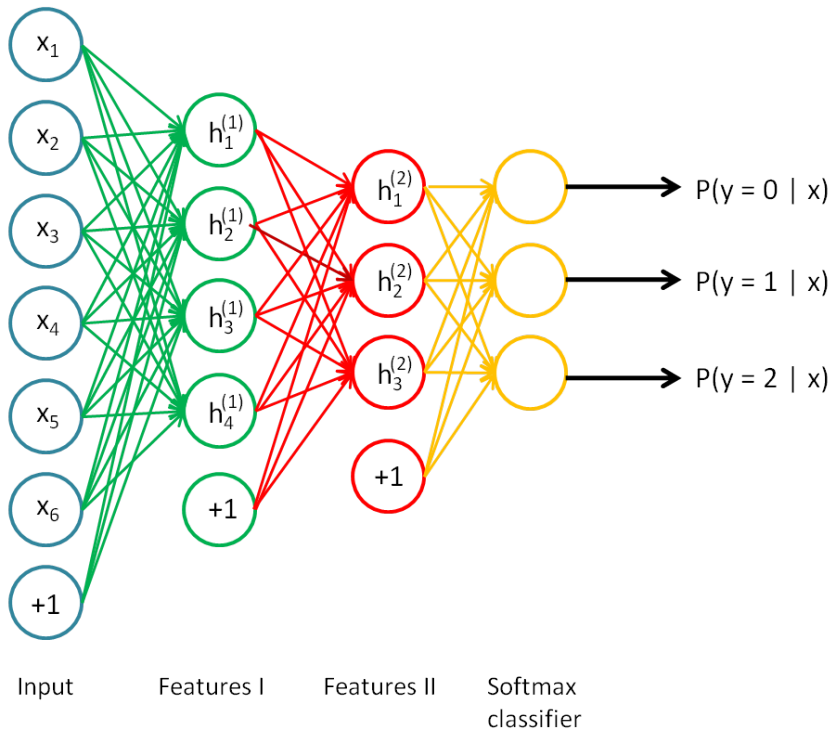


Figure: Original MJF dataset using echonest features visualized by t-SNE

Deep Learning

Deep neural network [LeCun et al., 1989]:

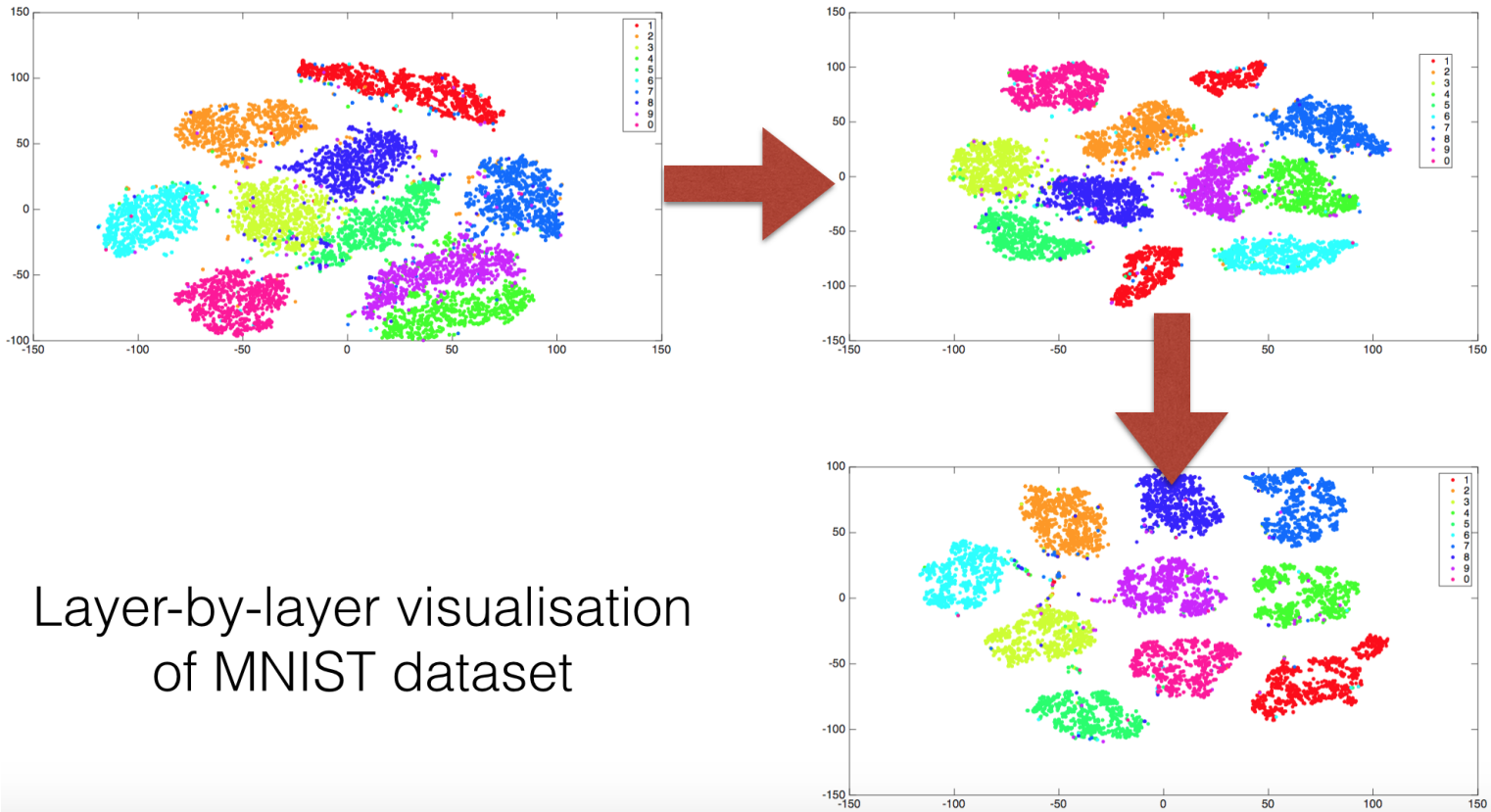
- uses a cascade of nonlinear layers for feature extraction and transformation.
- is based on the unsupervised learning of multiple levels of features of the data. Higher level features are derived from lower level features to form a hierarchical representation.



Deep Learning

Assumption

Deep Learning should distangle complex and high-dimensional datasets by learning new representations in which classes are more and more discriminated.



Layer-by-layer visualisation of MNIST dataset

Visualisation of GTZAN ([Sturm '13] Features)

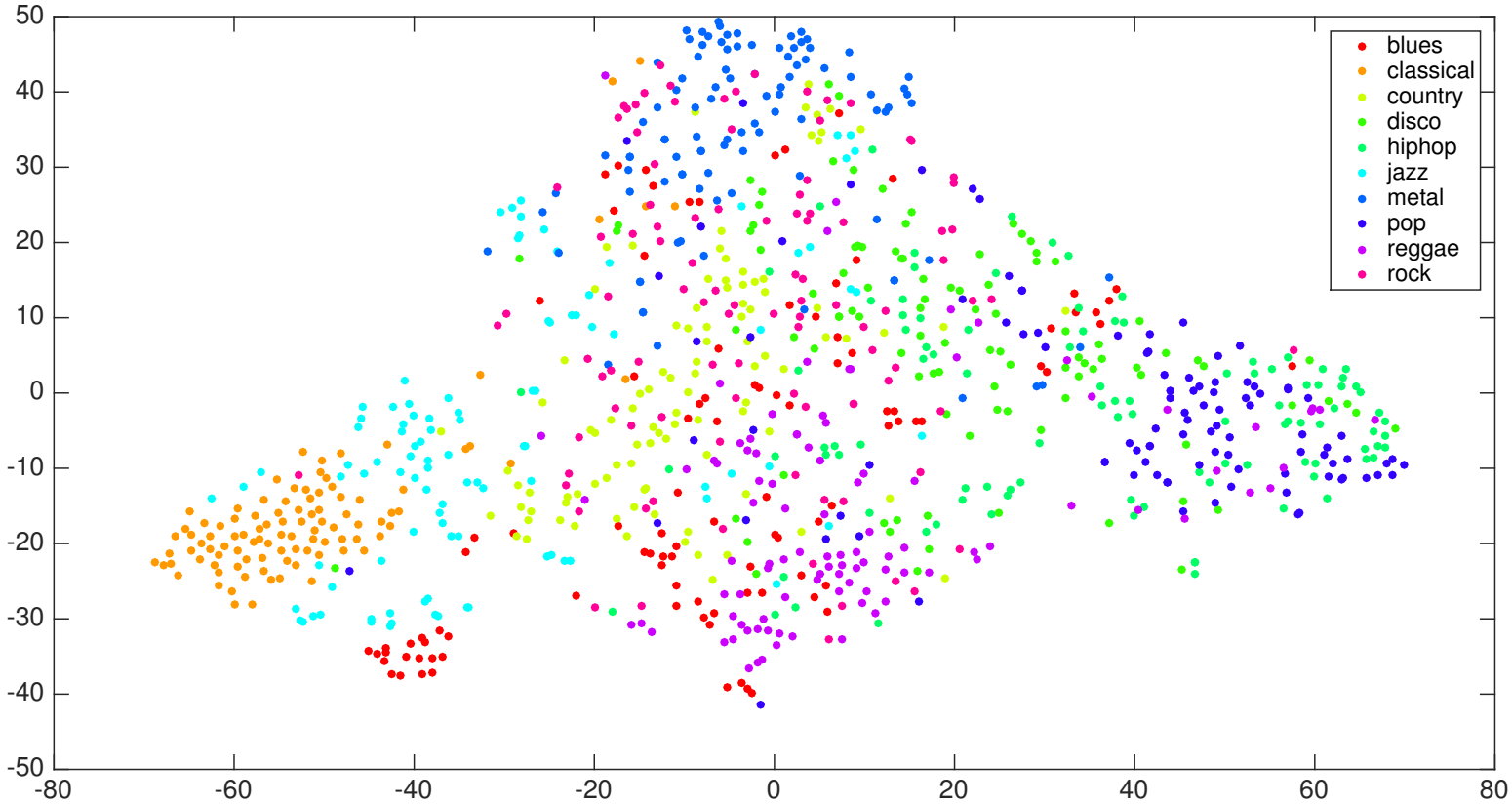


Figure: Original GTZAN dataset using [Sturm '13] features visualized by t-SNE

Visualisation of GTZAN ([Sturm '13] Features)

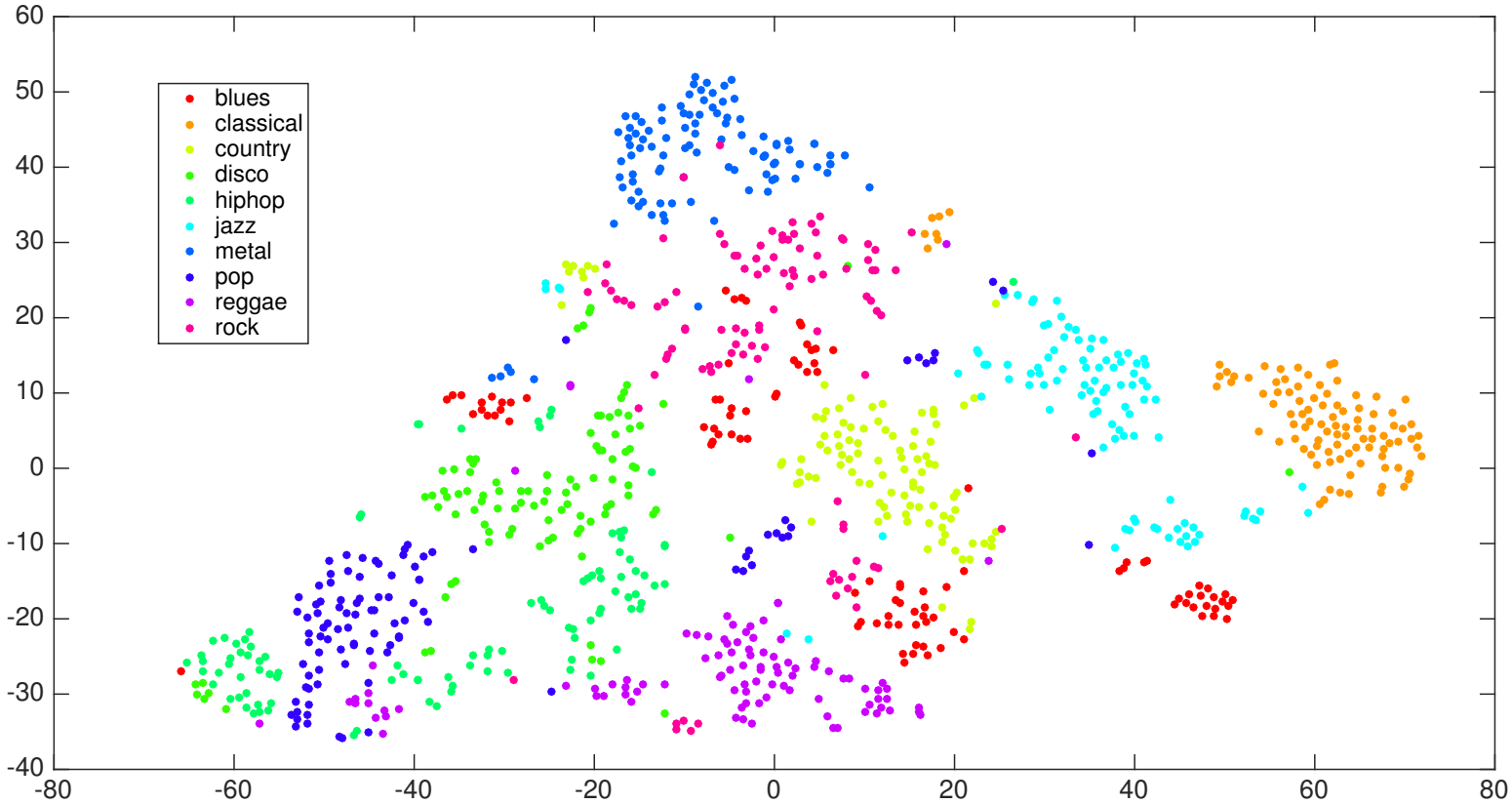


Figure: First-layer features of GTZAN dataset

Visualisation of GTZAN ([Sturm '13] Features)

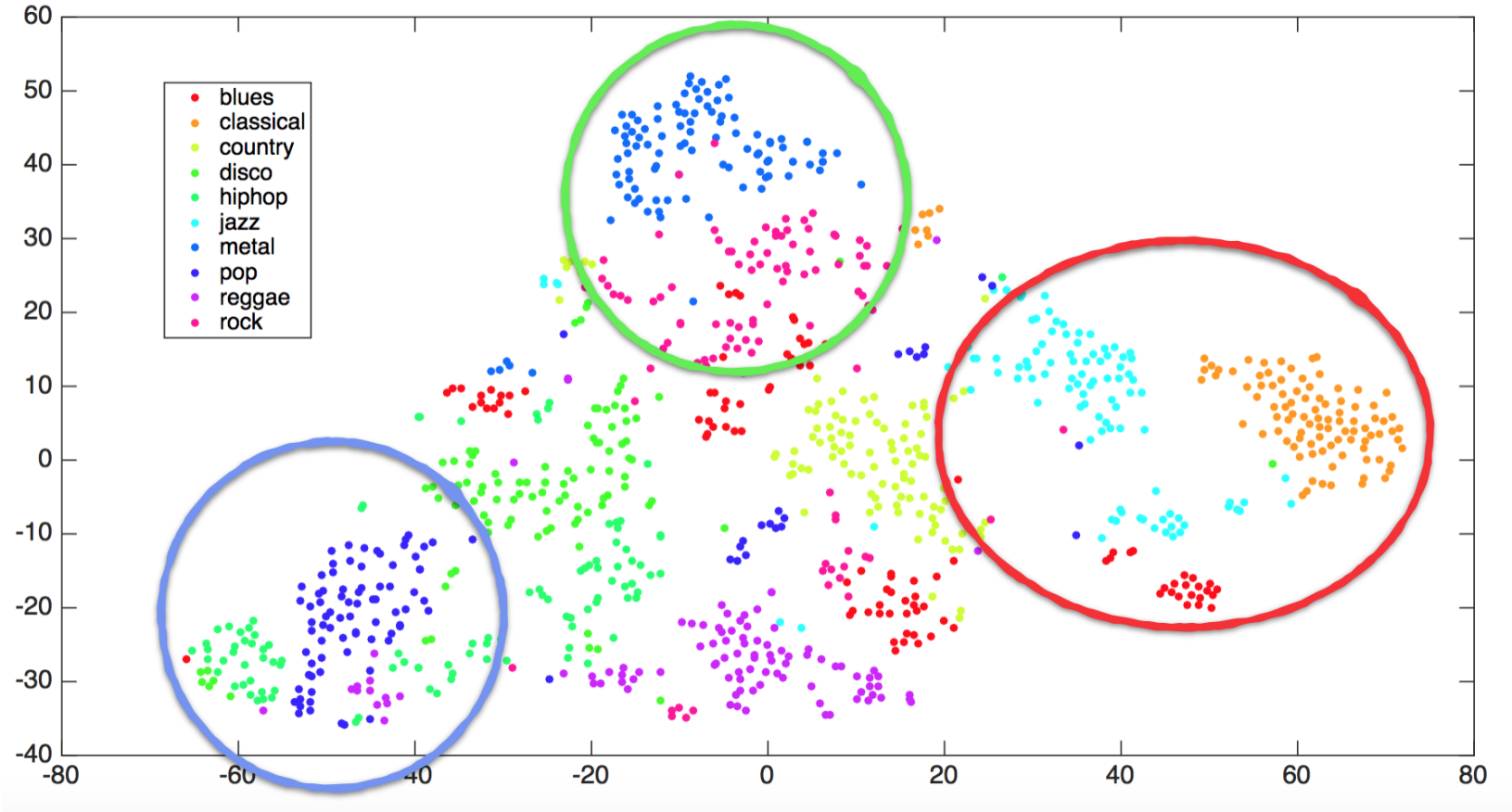


Figure: First-layer features of GTZAN dataset

Visualisation of GTZAN ([Sturm '13] Features)

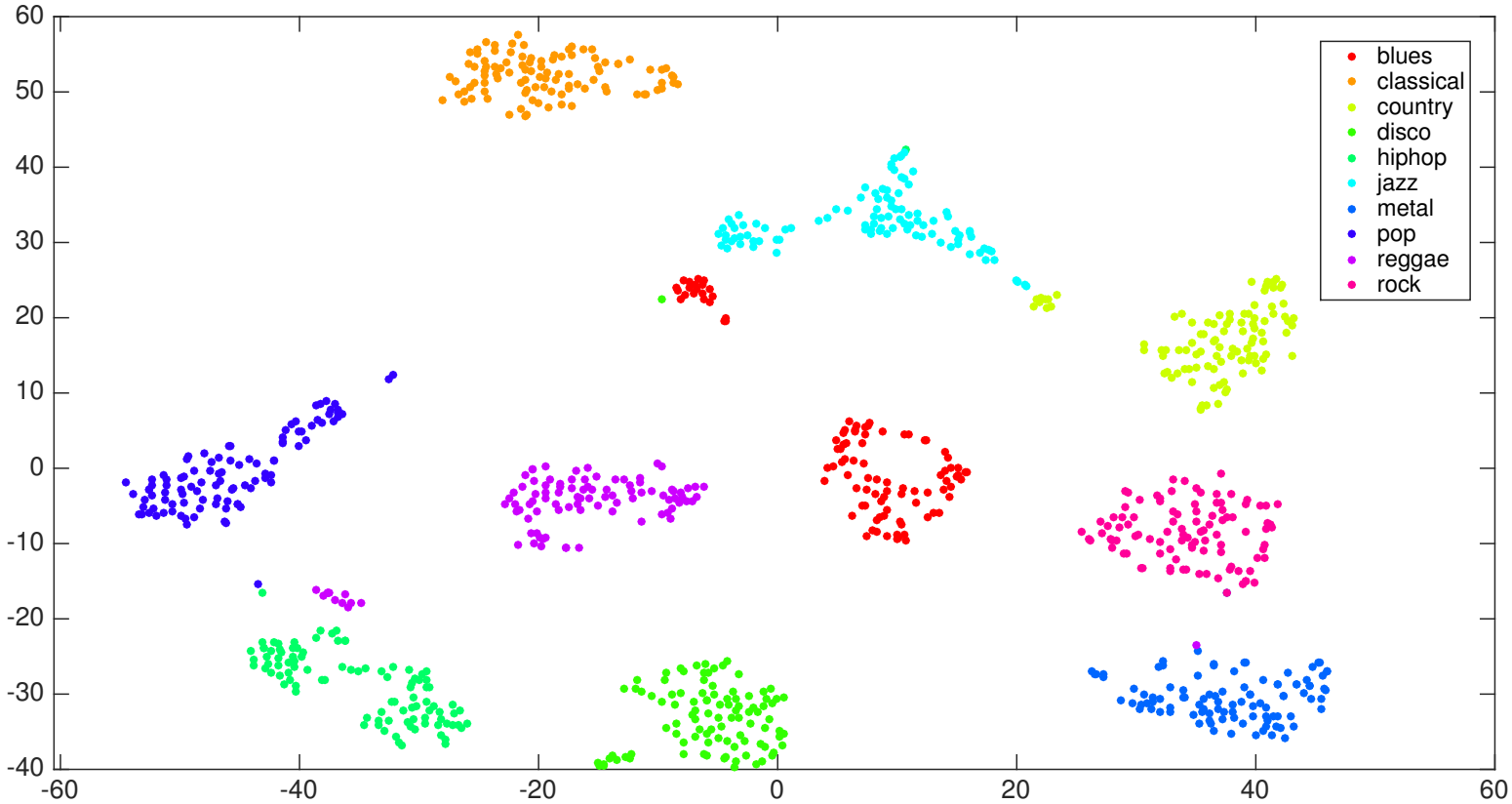


Figure: Second-layer features of GTZAN dataset

Visualisation of GTZAN ([Sturm '13] Features)

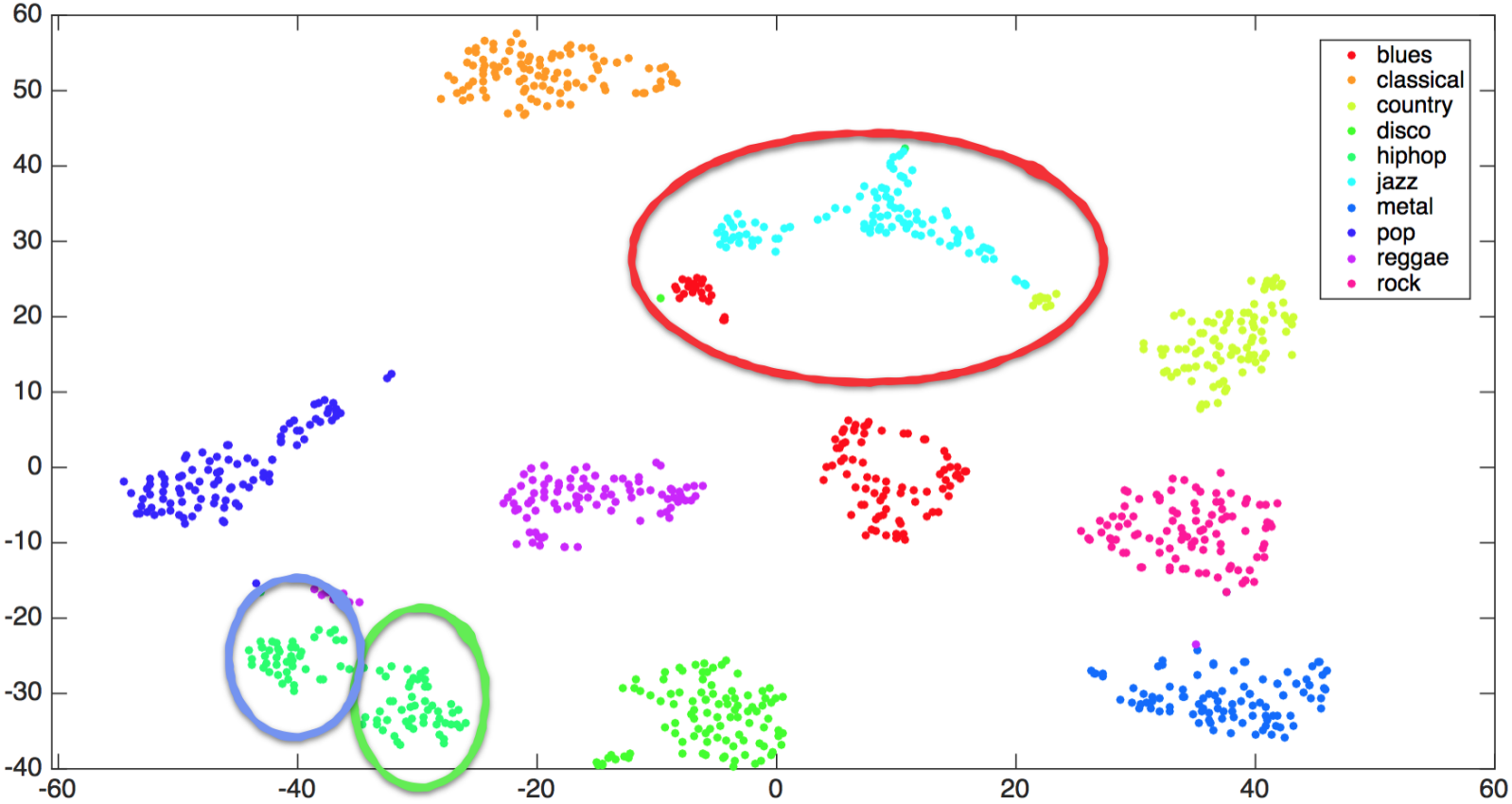


Figure: Second-layer features of GTZAN dataset

Visualisation of GTZAN (Learned Features)

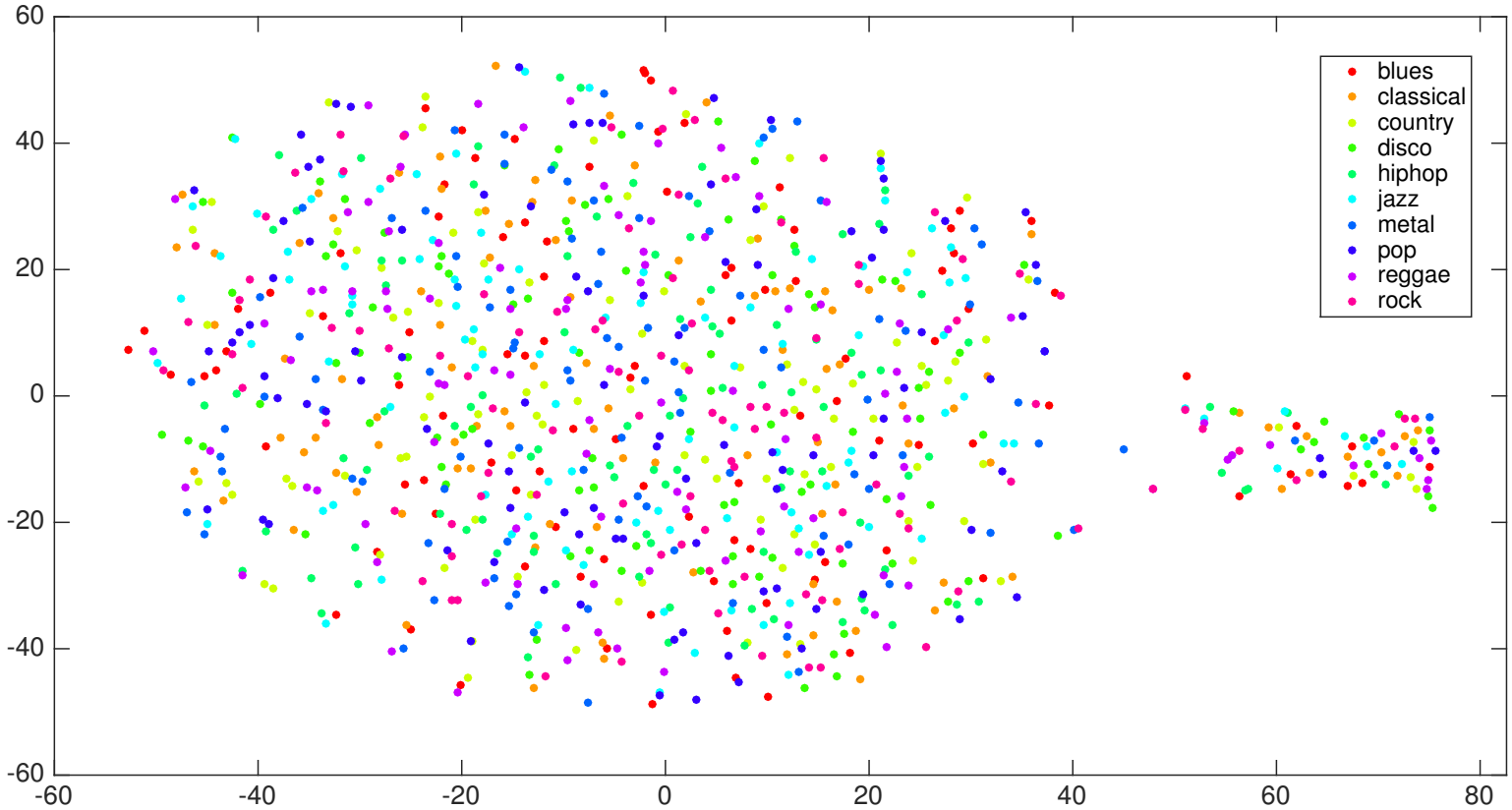


Figure: Original GTZAN dataset using learned features visualized by t-SNE

Visualisation of GTZAN (Learned Features)

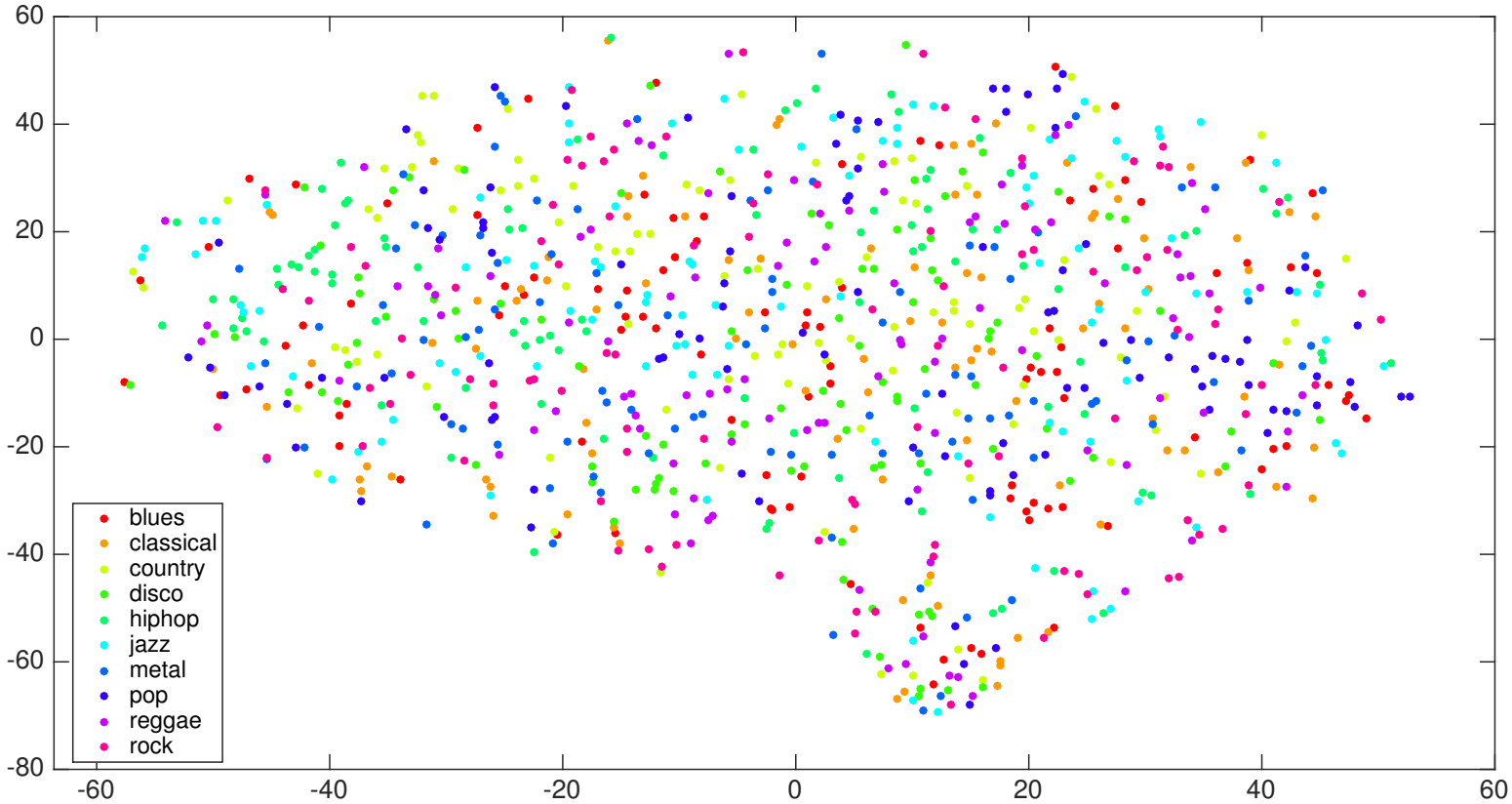


Figure: First-layer features of GTZAN dataset

Visualisation of GTZAN (Learned Features)

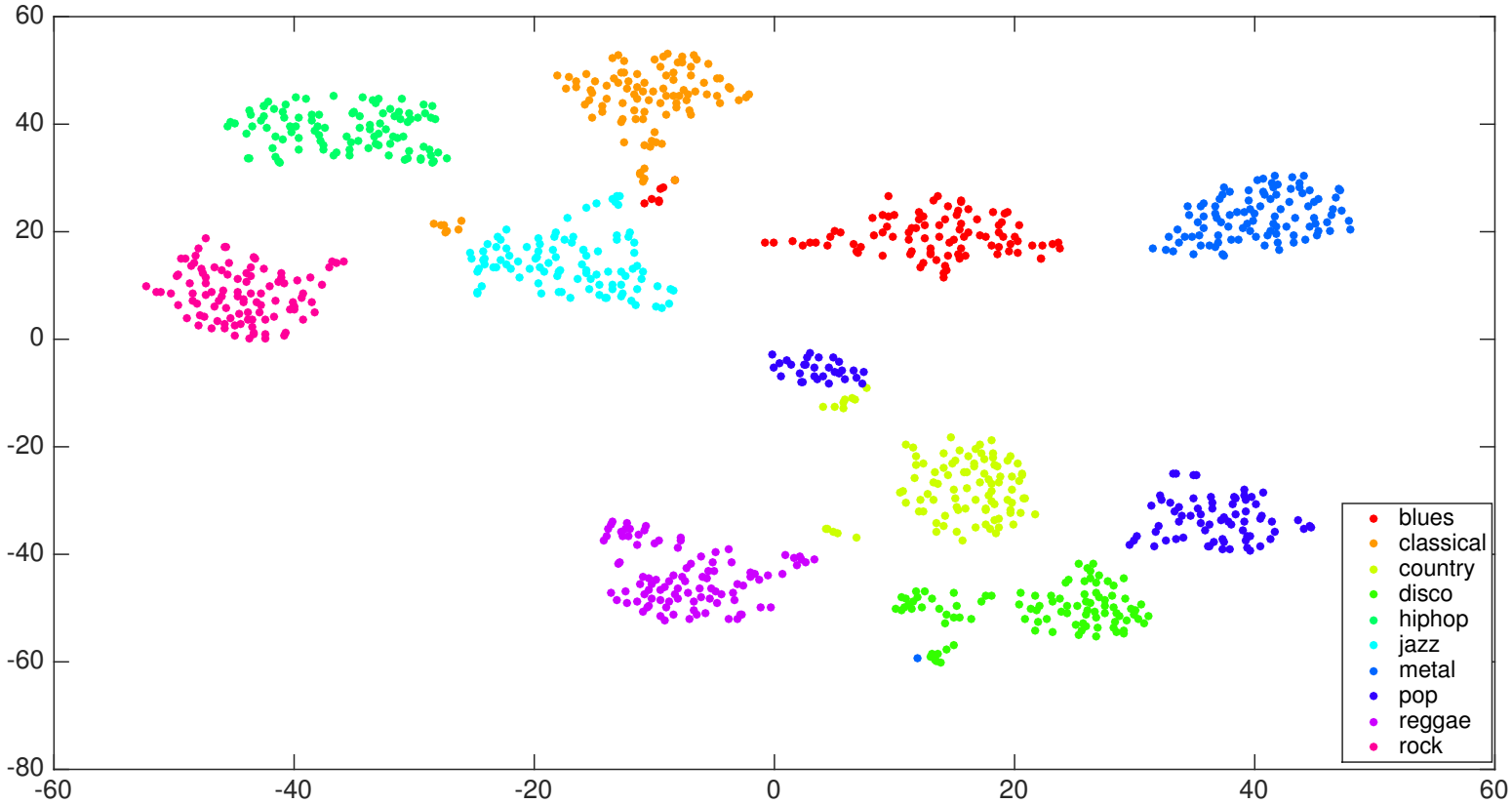


Figure: Second-layer features of GTZAN dataset

Visualisation of MJF Dataset

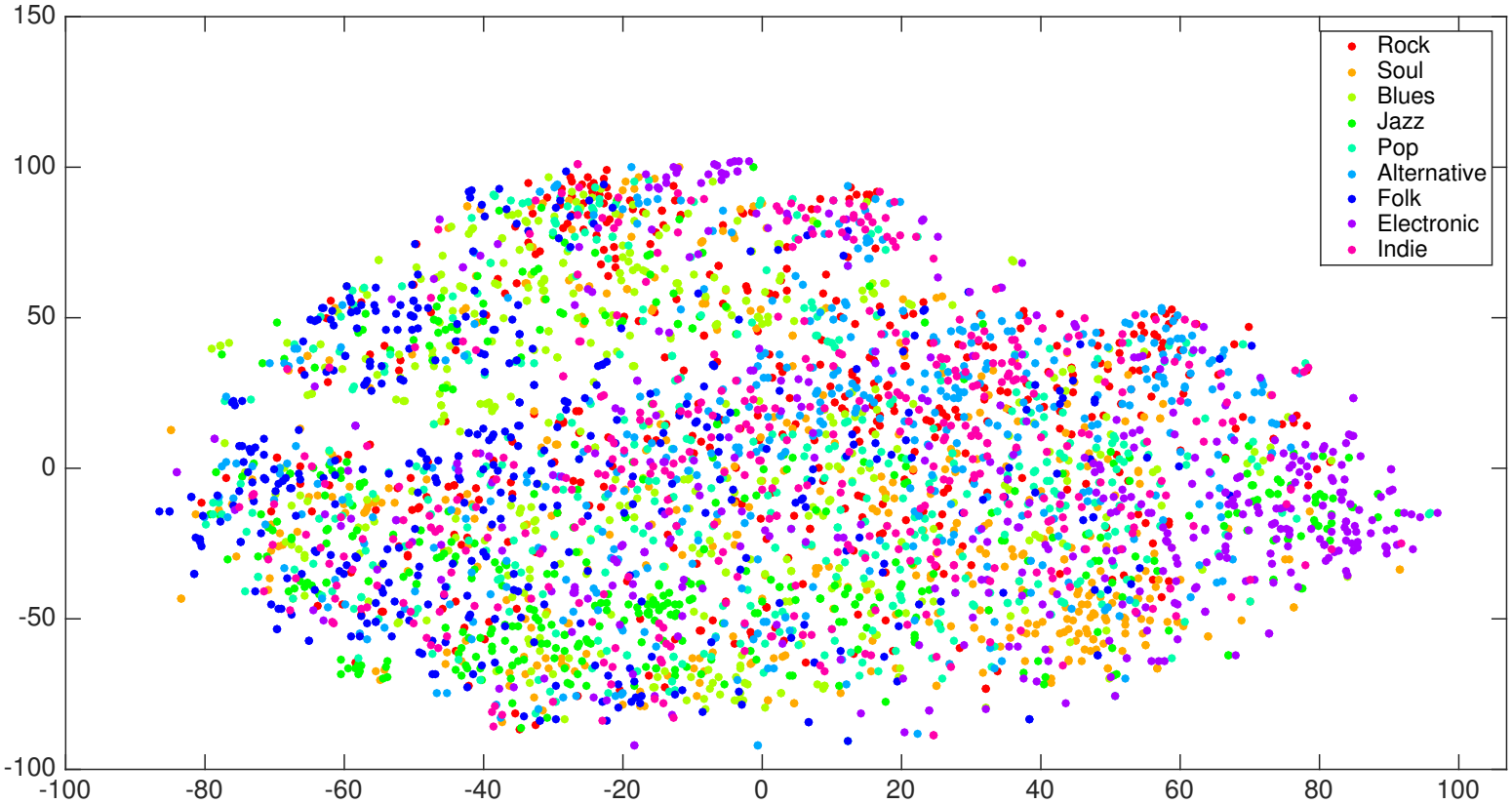


Figure: Original MJF dataset using echonest features visualized by t-SNE

Visualisation of MJF Dataset

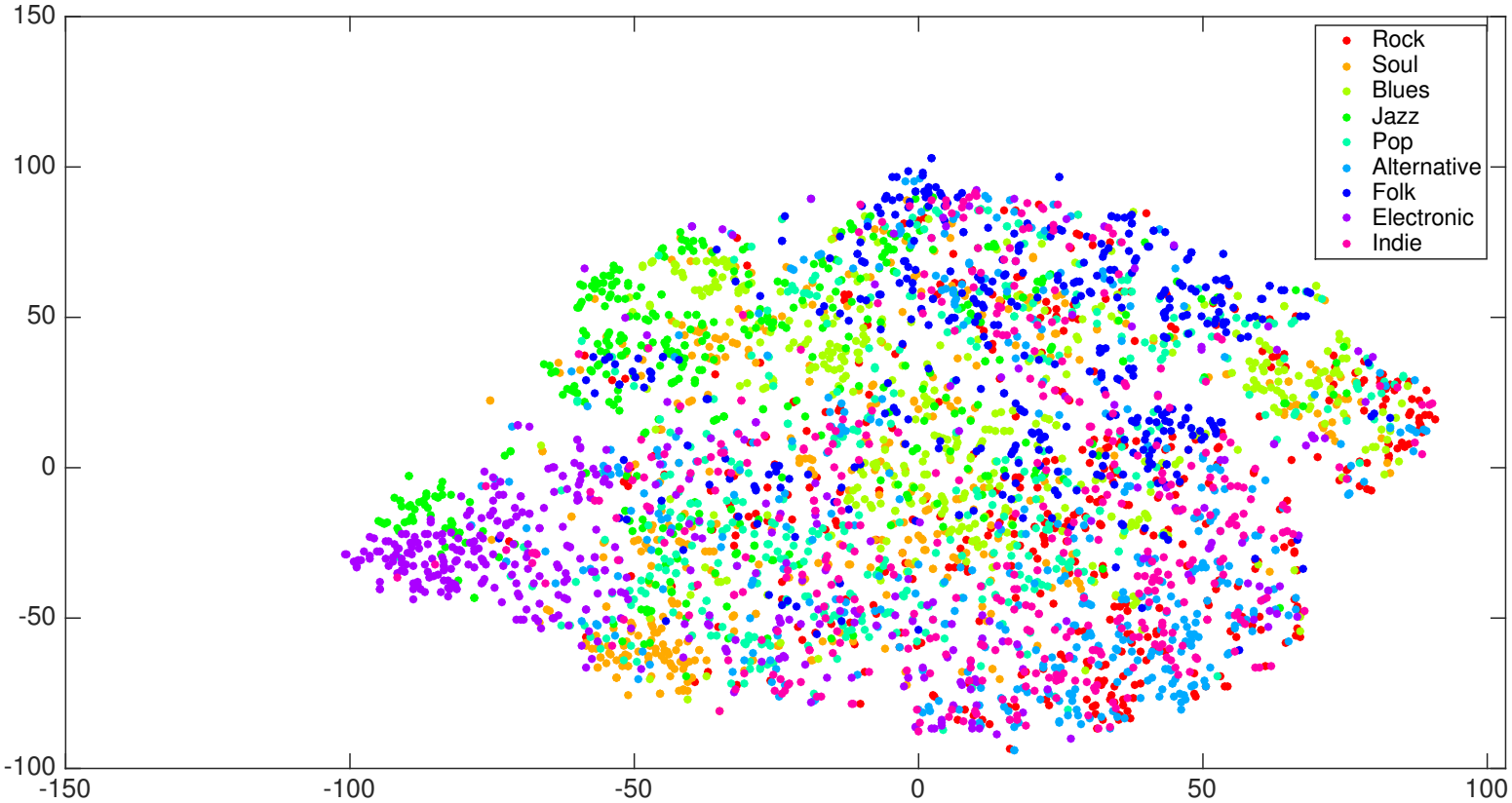


Figure: First-layer features of MJF dataset

Visualisation of MJF Dataset

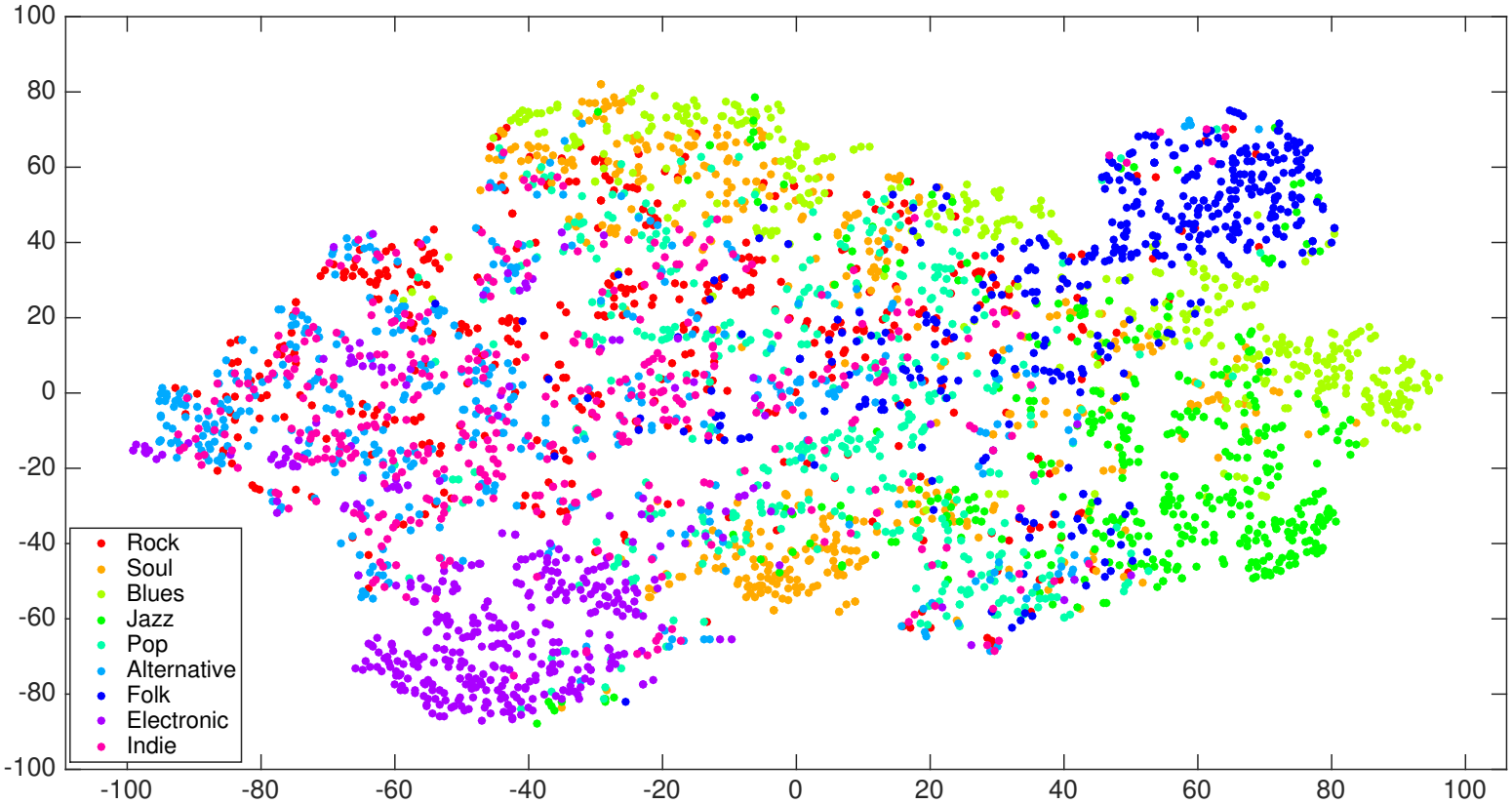


Figure: Second-layer features of MJF dataset

Visualisation of MJF Dataset

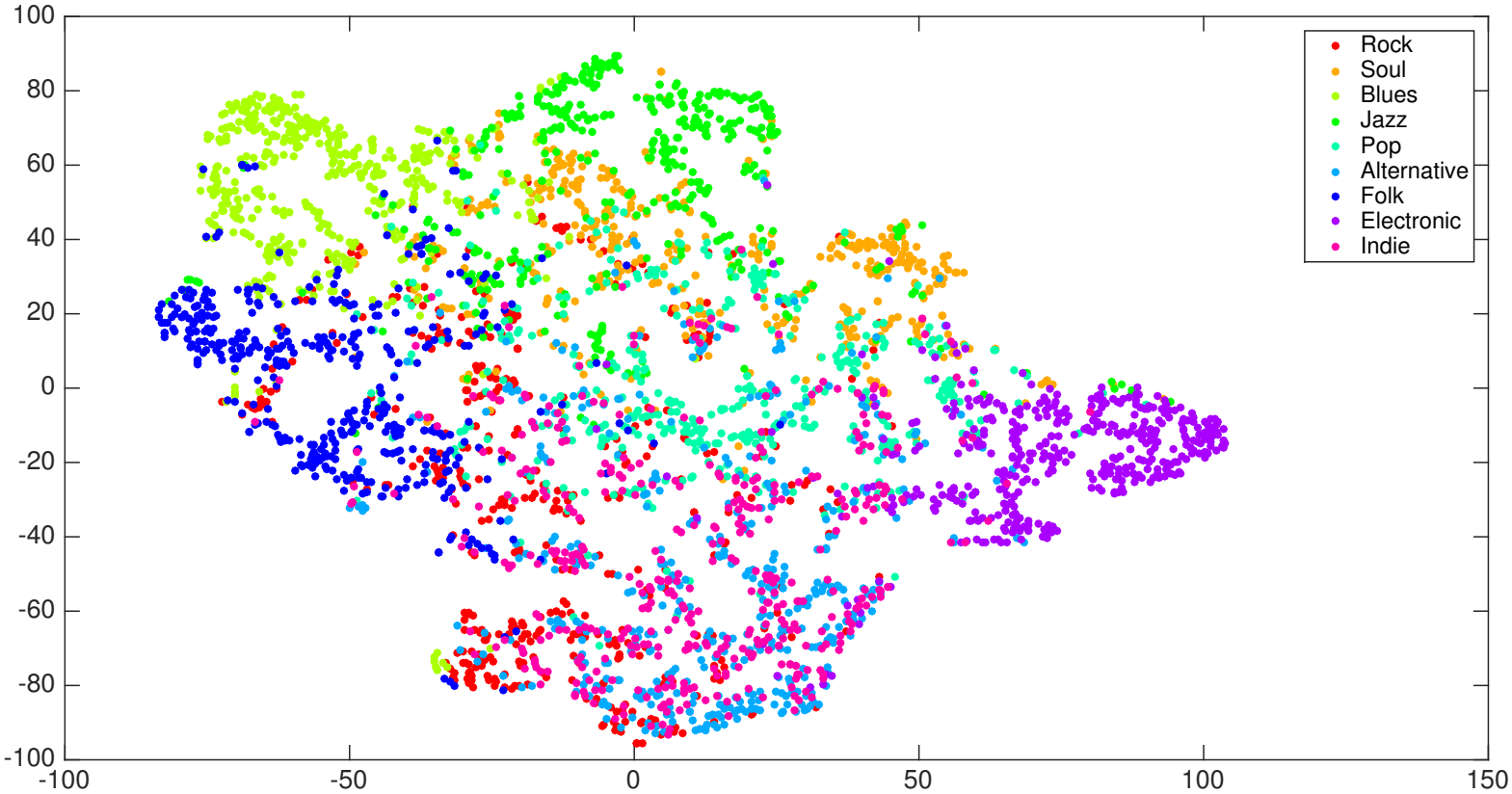


Figure: Third-layer features of MJF dataset

Visualisation of MJF Dataset

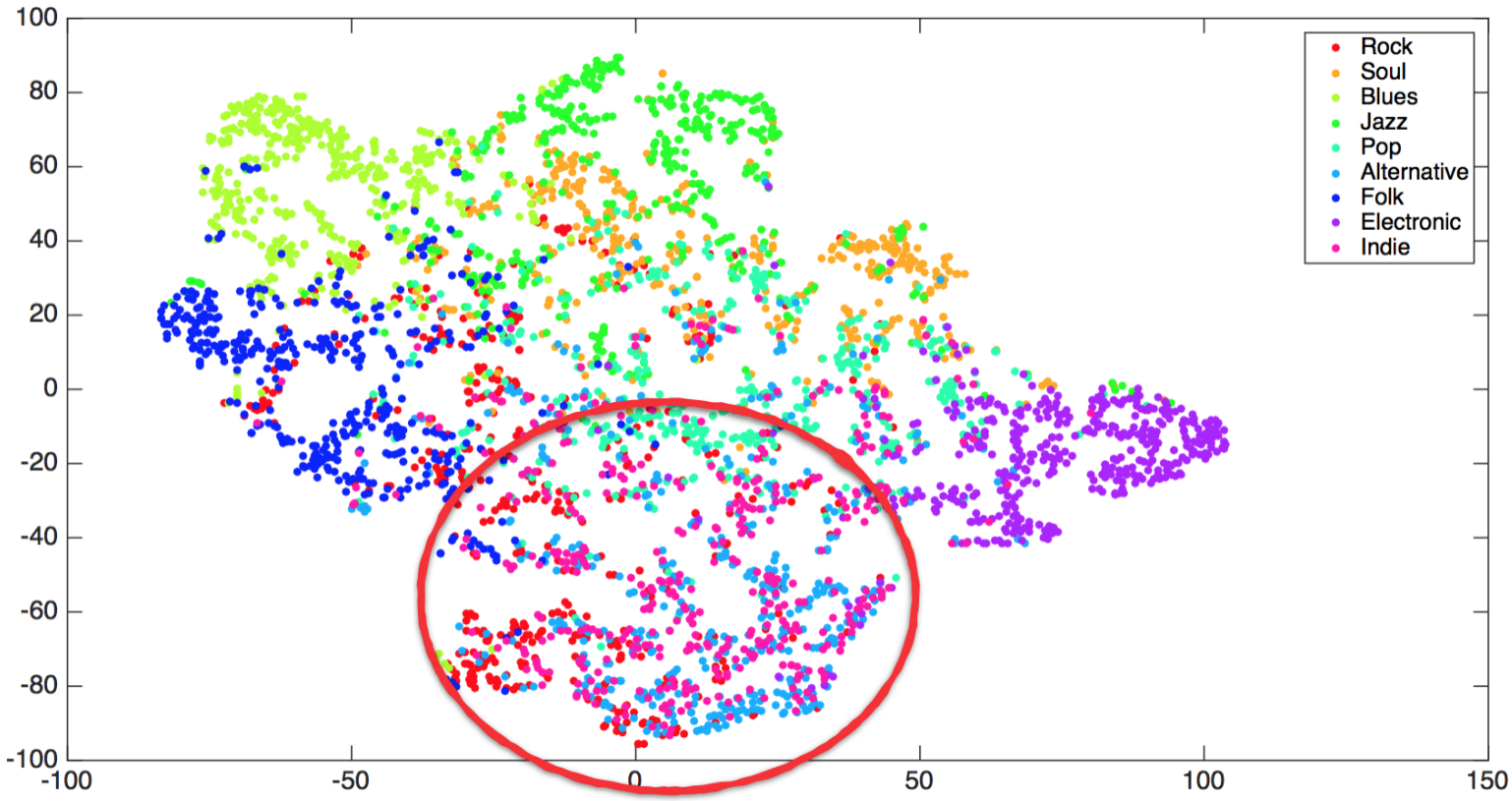


Figure: Third-layer features of MJF dataset

Visualisation of MJF Dataset

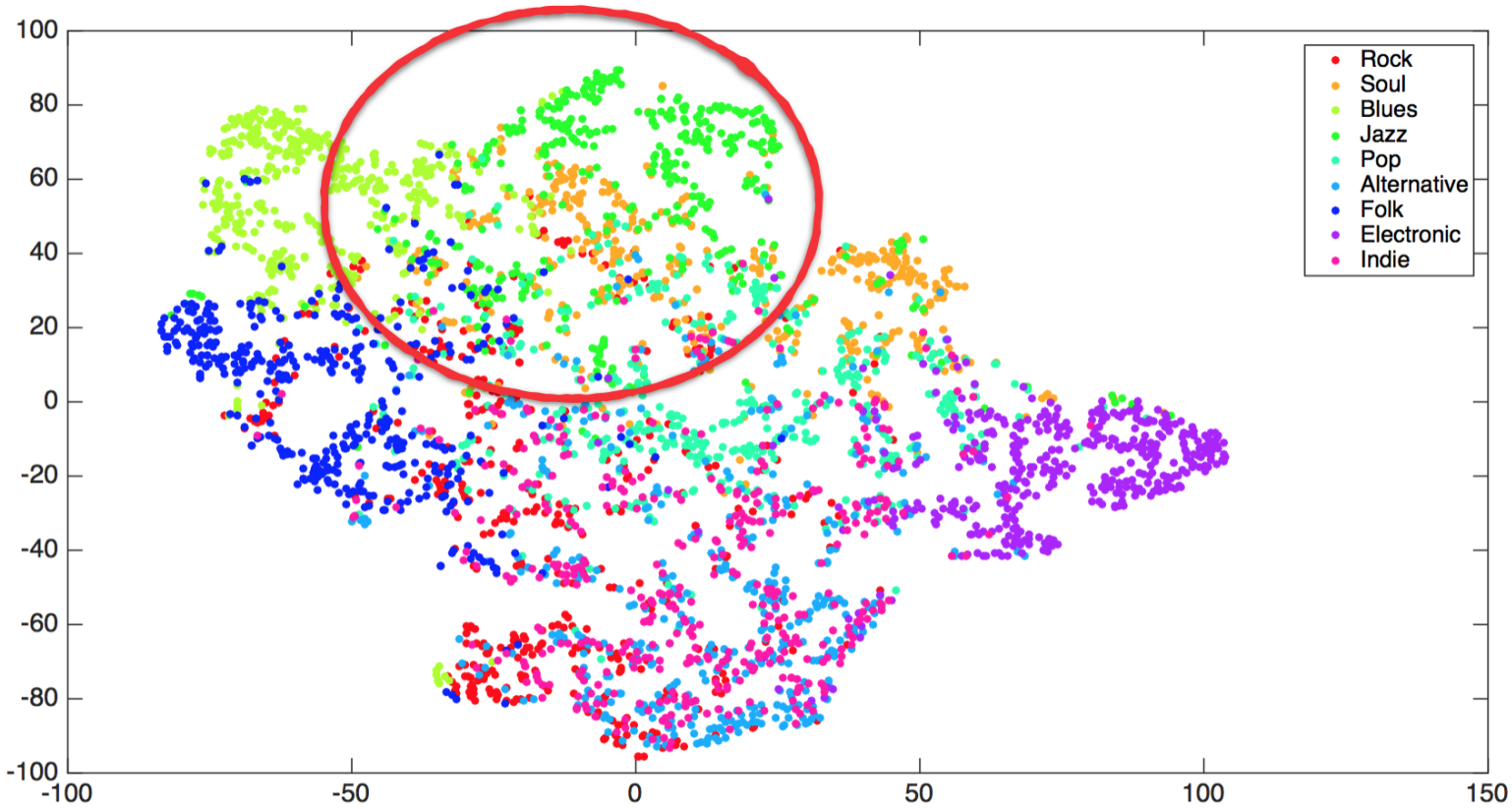


Figure: Third-layer features of MJF dataset

Music Recommendation - Shortest Path

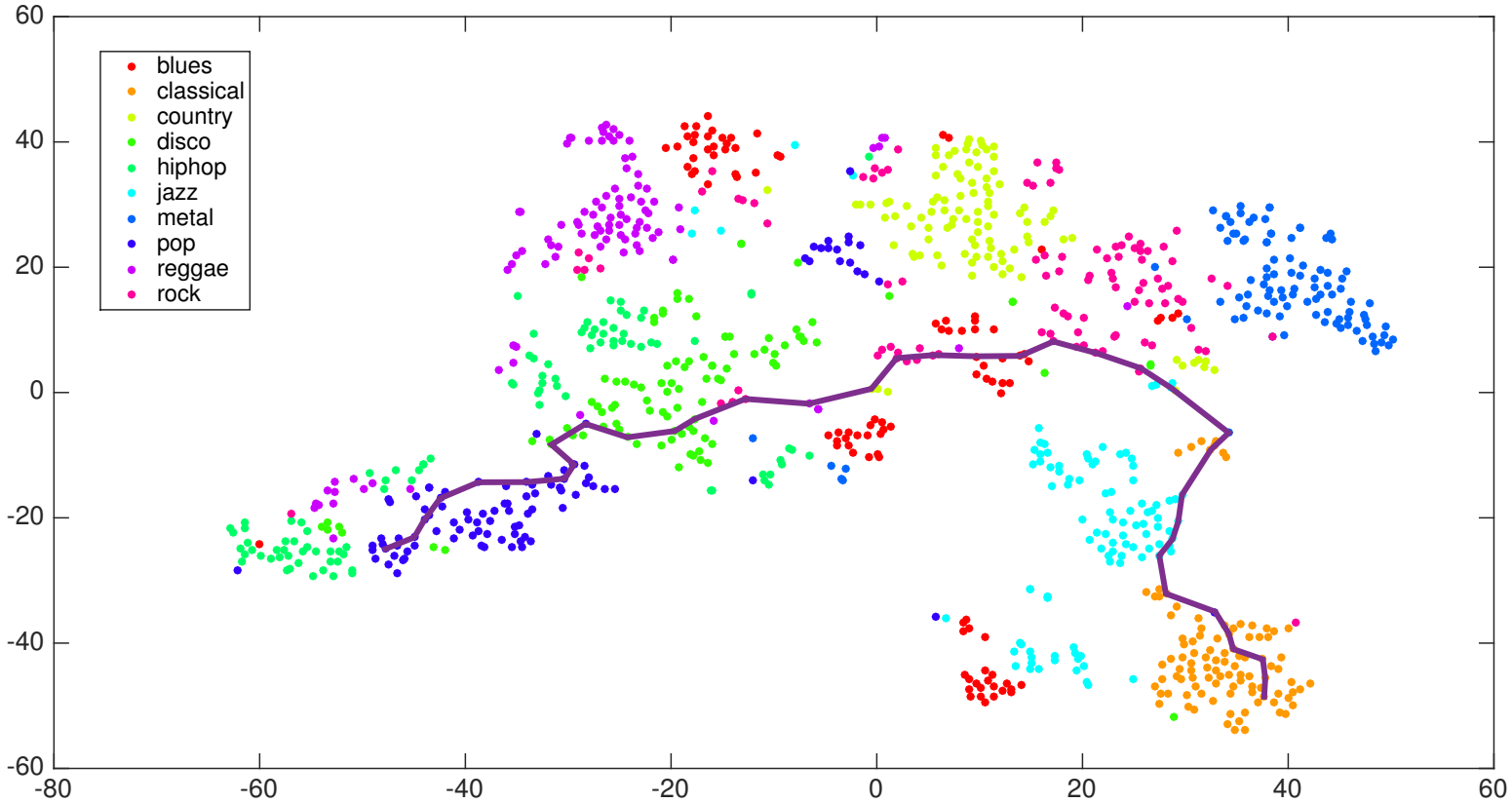


Figure: Playlist created by shortest path algorithm

Music Recommendation - Random Walk

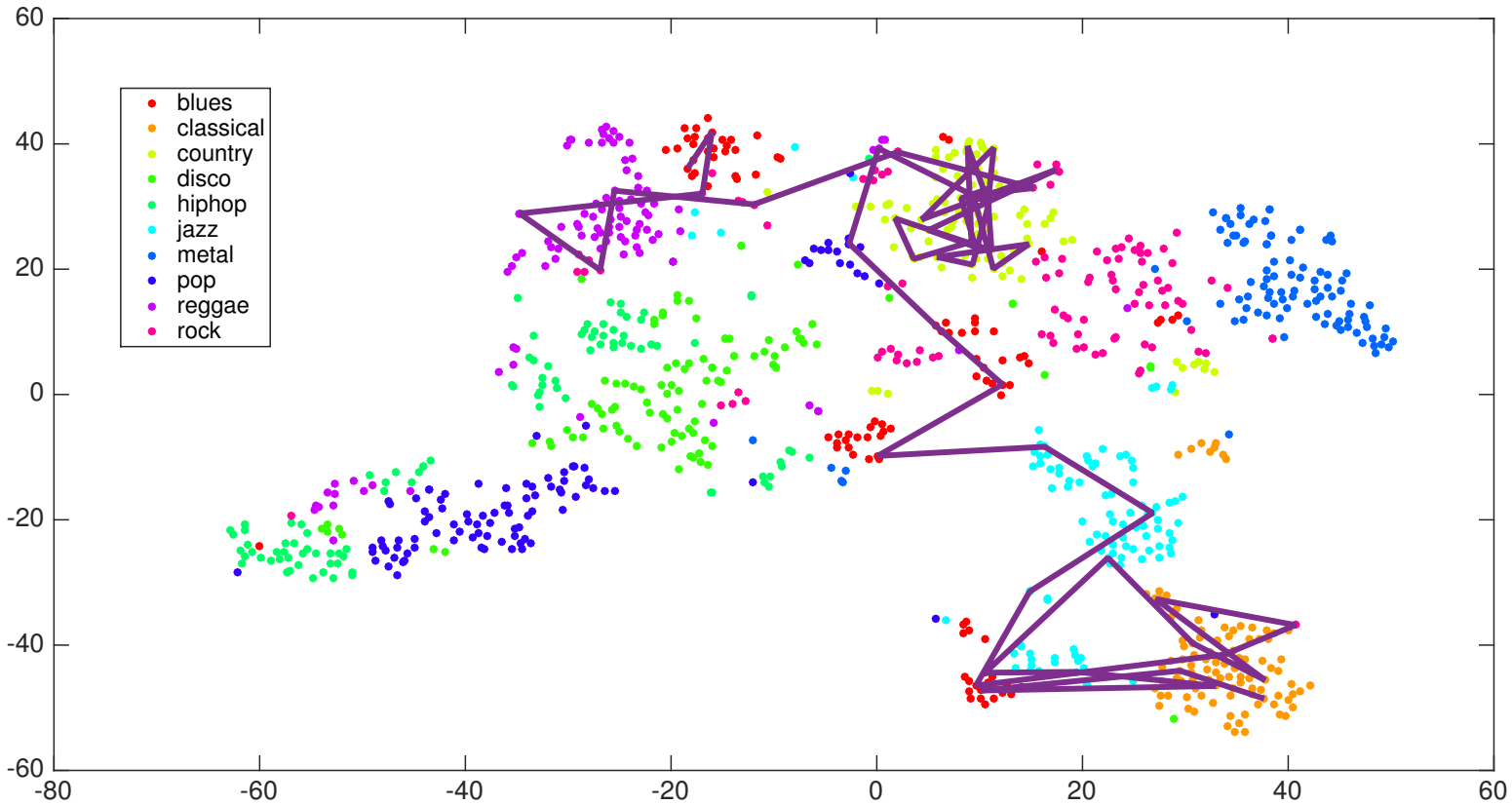


Figure: Playlist created by random walk algorithm

Conclusion

- Deep Learning + t-SNE is a promising visualisation tool for complex high-dimensional datasets.
- Applications - music recommendation and structure mining.
- Future development - to combine visualisation with Deep Learning.

References

1 Sturm, B.L., "On music genre classification via compressive sampling," Multimedia and Expo (ICME), 2013 IEEE International Conference, vol., nn., pp.1,6, 15-19 July 2013

2 Tristan Jehan, T. and DesRoches, D., "Analyzer documentation", 2014. Available online at `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`; visited on May 26th 2015

3 Montreux Jazz Festival - `http://www.montreuxjazz.com/`

4 LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Computation, 1, pp. 541–551, 1989.

References

5 Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals," IEEE Transactions on Audio and Speech Processing, 2002

6 Schindler, A. and Rauber, A., "Capturing the temporal domain in echonest features for improved classification effectiveness," Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation, pp. 214–227. Springer, 2014

7 Van der Maaten, L.J.P. and Hinton, G.E., "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research 9(Nov), pp. 2579-2605, 2008.

8 Henaff M, et. al., "Unsupervised learning of sparse features for scalable audio classification," ISMIR, 2011.