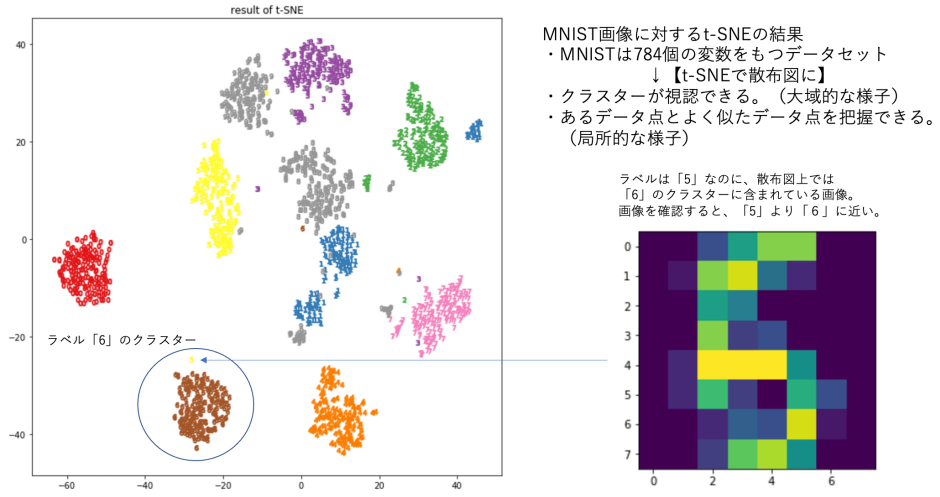
t-SNE・UMAPと次元削減

1.　t-SNE・UMAPとは？

手元にd個の変数からなるデータセット$\mathcal{D}$があるとします。dが大きいと、pairplot（2変数どうしの散布図）などで可視化しても、データセットの様子を把握することが難しくなります。そこで、データセット$\mathcal{D}$のd次元散布図における

- 大域的な様子：含まれているクラスターはなにか。
- 局所的な様子：あるデータ点の周辺にある他のデータ点はなにか。

を多くて3次元程度の散布図Sで視認できるようにならないでしょうか。



この課題を解くために提案された手法が**t-SNE**（t-distributed Stochastic Neighbor Embedding, t-分布型確率近傍埋め込み）と**UMAP**（Uniform Manifold Approximation and Projection）です。

手法	提案された年	著者
t-SNE	2008	L. van der Maaten, G. Hinton
UMAP	2018	L. McInnes, J. Healy, J. Melville

2.　t-SNE・UMAPの仕組み

t-SNE・UMAPの前身といえる手法に多次元尺度構成法（MultiDimensional Scaling, MDS）があります。まずMDSの仕組みを説明し、次にMDSと比較する形でt-SNEとUMAPの仕組みを解説し、最後にt-SNE・UMAPがMDSに比べてどのような特徴を持っているのか解説します。