UMAP (Uniform Manifold Approximation and Projection) est une technique de réduction de dimension qui peut être utilisée pour la visualisation de données de manière similaire à la t-SNE. Cette technique peut aussi être utilisée pour de la réduction de dimension non linéaire. L’algorithme de cette technique est basé sur 3 hypothèses faites sur les données :

1. Les données sont distribuées de manière uniforme sur une variété Riemannienne
2. La métrique de Riemann est localement constante (ou peut être approximée comme tel)
3. Les « points » sont localement connecté dans l’espace, il n’y a pas de « points » isolés

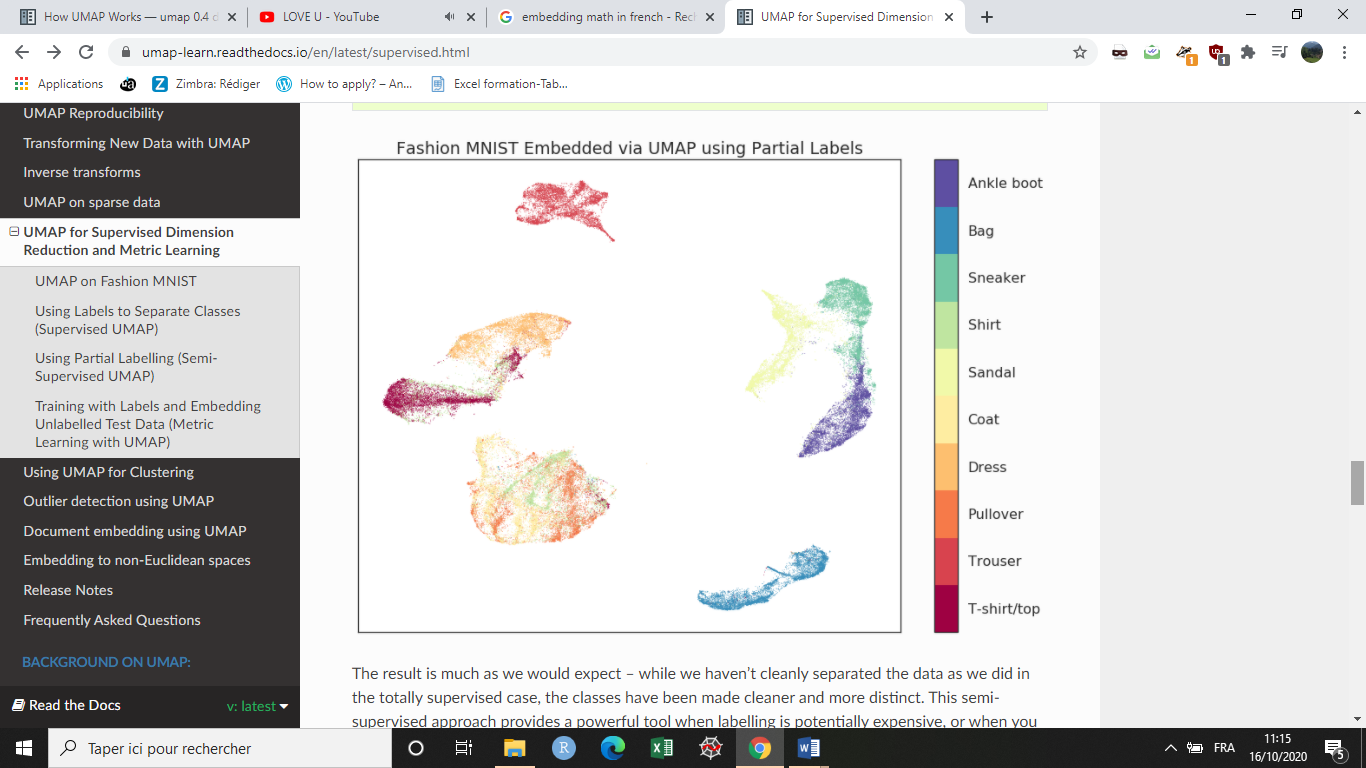
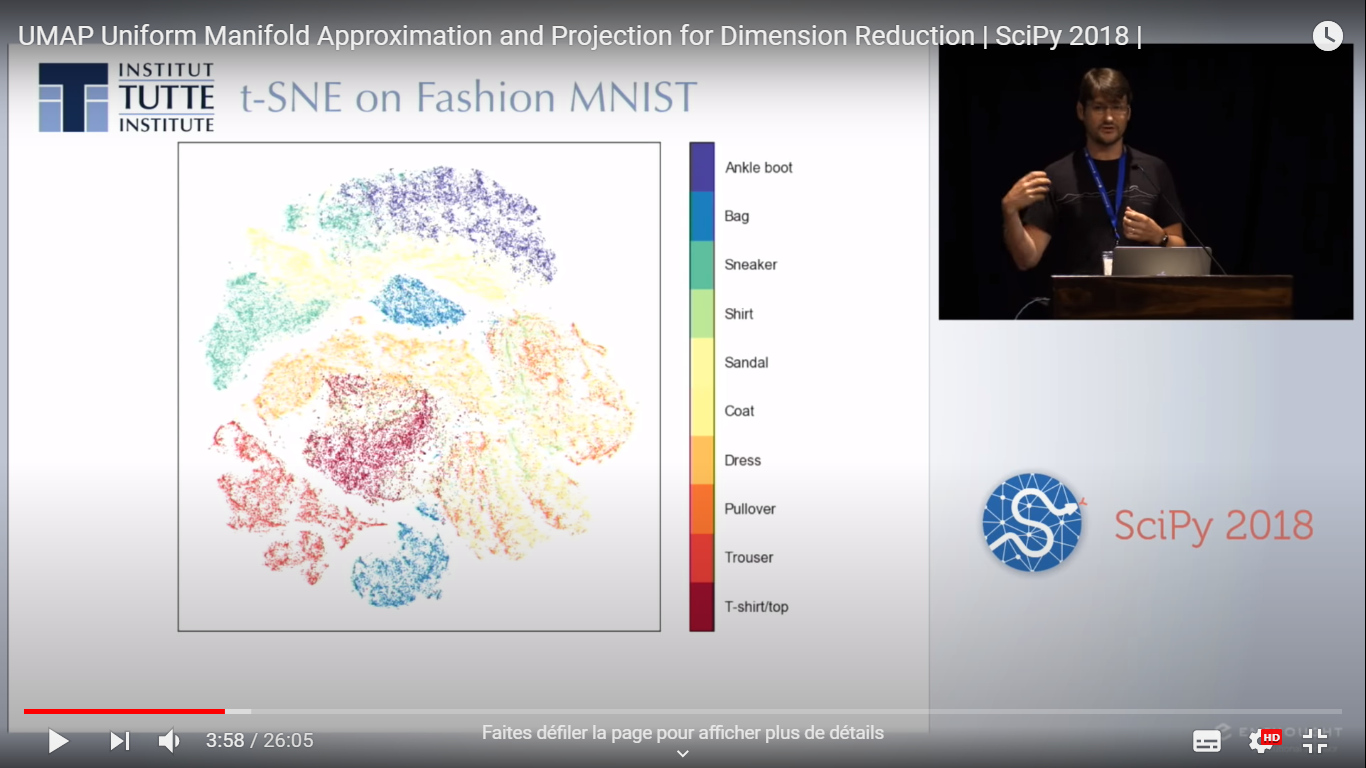
De ces 3 hypothèses il est possible de modéliser les données avec une structure topologique floue (fuzzy topological structure) (avant projection). La représentation (simple) des données est trouvée en cherchant une projection (des données) sur un espace de faible dimension qui a une structure la plus proche possible de la structure topologique de départ.

Domaines mathématiques derrière la méthode : algèbre topologique, Géométrie de Riemann, « logique flou »

Video (<https://www.youtube.com/watch?v=nq6iPZVUxZU>):

UMAP est basée sur des techniques de Graphe de voisins (Neighbour Graphs Techniques). L’état de l’art de ces techniques étant la méthode t-SNE. Cependant la méthode UMAP ne se limite pas à capturer la structure locale des données mais également la structure globale.

Exemple :

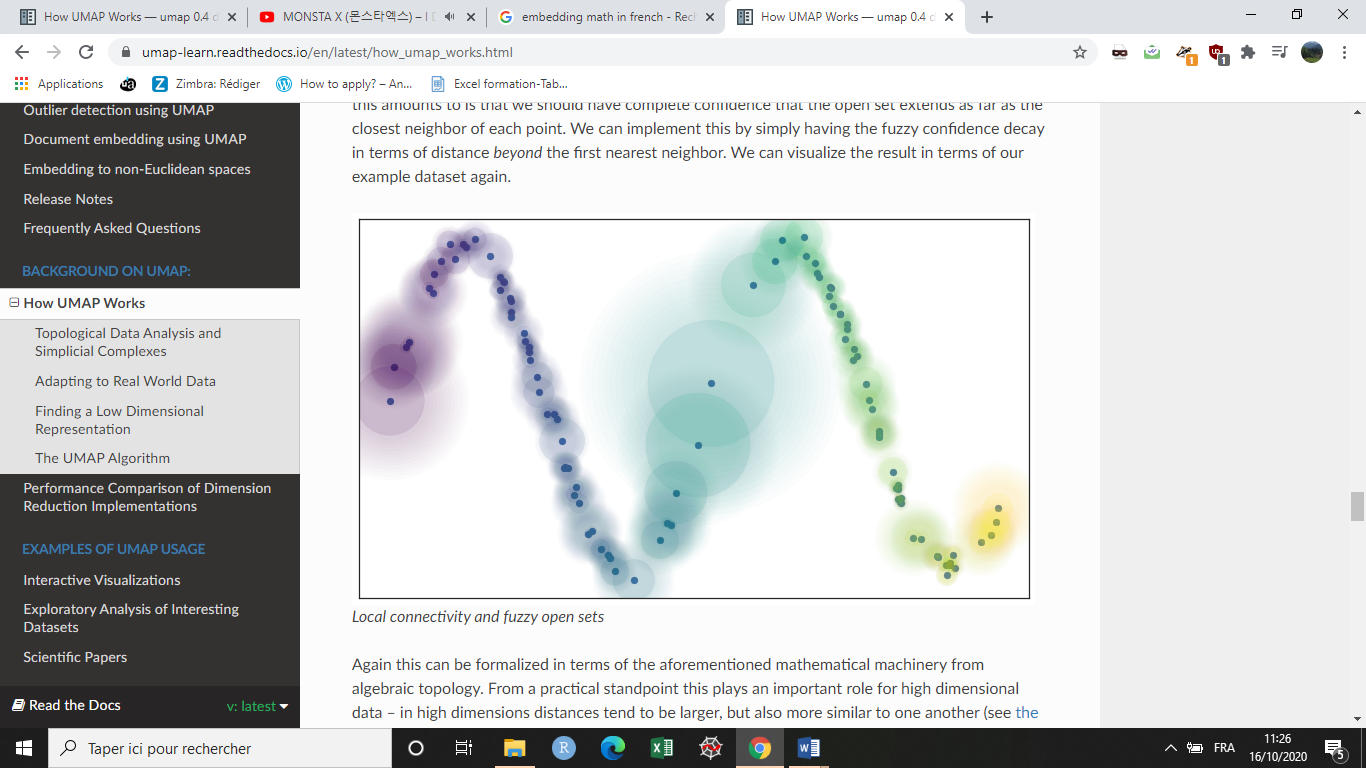
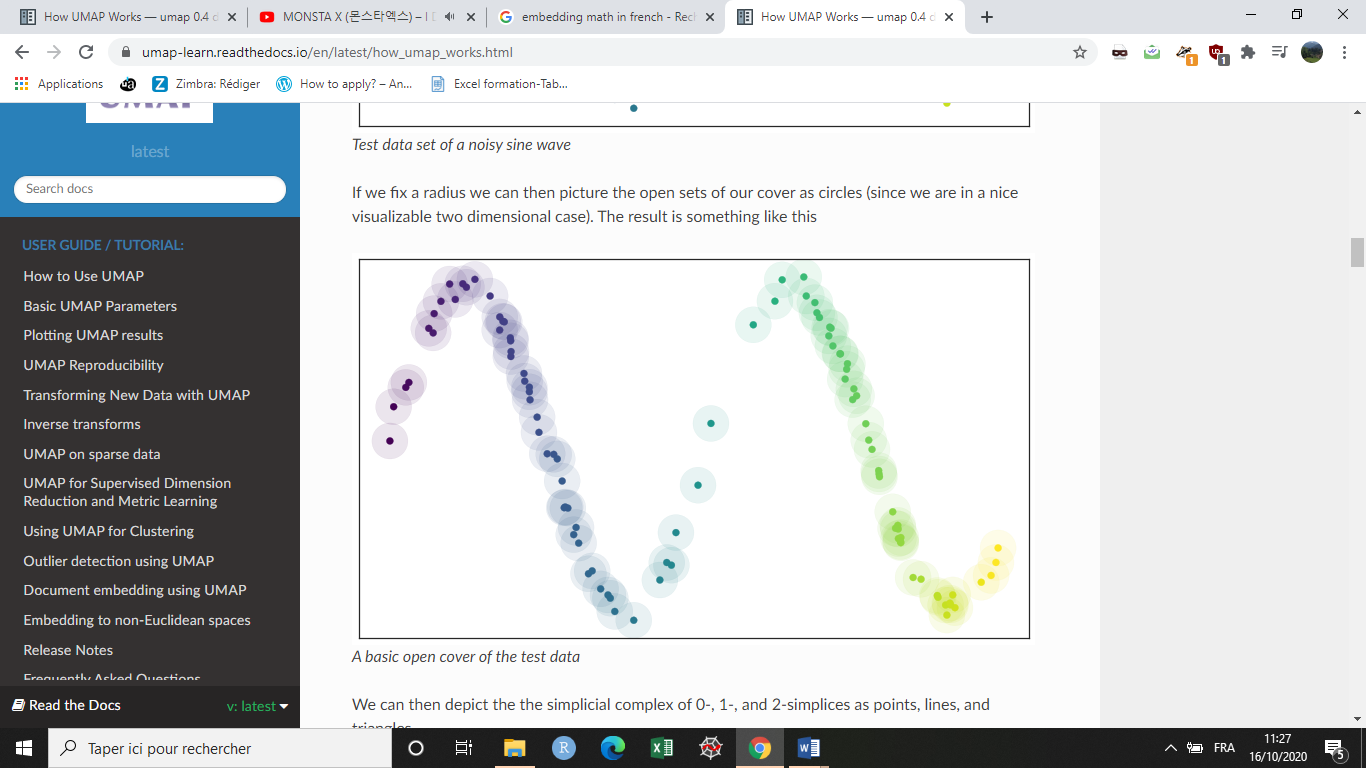
 

On voit que les 2 techniques on réussies à séparer les groupes de manière distincte mais la technique UMAP également réussi à capter la structure des globales des données en rapprochant dans l’espace les bottes et les baskets par exemple, en les éloignant les T-shirts. (Données : fashion MNIST)

Au niveau mathématique :

1. Analyse de données topologiques : Il est possible en construisant d’une certaine manière des   
   complexes simpliciaux dans un espace topologique de les reconstruire de manière combinatoire sans perdre d’information (on réussit à recouvrir toute l’information importante sur la topologie de l’espace de départ). Ce qui est plus simple à manipuler.
2. Hypothèse de distribution uniforme : Si les données ne sont pas uniformément distribuées sur la variété, on peut définir une métrique Riemannienne pour faire en sorte que l’hypothèse soit vérifiée, en faisant varier la notion de distance pour chaque type de données.

Exemple :

A droite, la même métrique a été appliquée pour tous les points. Toutes les sphères sont des sphères unités. A gauche, on a appliqué des métriques différentes. Les sphères sont également toutes des sphères unités mais en considérant chacune sa propre métrique.

Au niveau algorithmique :

1. Quand on connait la dimension de l’espace sur lequel on projette les données (typiquement R² ou R3), on ne connait alors plus la bonne distance des voisions les plus proches. On a alors besoin de faire appel à l’entropie croisée pour déterminer les hyper paramètres. (Principe de l’entropie croisée : le premier terme permet de récupérer la structure locale des données (classes/groupes), le deuxième terme permet de récupérer la structure globale des données (écart entre les différentes classes/groupes)). Concrètement cela est implémenté à l’aide des RP-trees (Rare Pattern Tree Mining) et NN-descent (Nearest Neighbor descent).
2. La rapidité de l’algorithme est assurée par SGD (Stochastique Gradient Descent) + negative sampling

Avantages :

1. Algorithme bien plus rapide que la t-SNE (écart de plus en plus grand avec la taille des données)
2. Peut prendre en charge plusieurs types de données en même temps, chacune ayant des métriques différentes.
3. Peut faire de la classification supervisée comme non supervisée.