REDUCTION DE DIMENSIONS

UMAP ET TMAP

REDUCTION DE DIMENSIONS

1. Qu'est-ce qu'une reduction de dimension ?
2. Pourquoi est-elle necessaire?
3. Résumé des techniques existantes
4. Explications UMAP et TMAP

1. QU'EST-CE QU'UNE REDUCTION DE DIMENSION ?

- En machine learning, la quantité d'information à traiter pour faire des predictions, classifications ou clustering est souvent énorme.
- Problème : Fléau de la dimension : le volume de l'espace croît rapidement si bien que les données se retrouvent « isolées » et deviennent éparses diminuant leur pertinence ou inefficaces.
- Solution : prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension

2 . POURQUOI EST-ELLE NECESSAIRE?

Raison #1

- Espace necessaire au stockage des données diminue
- Moins de temps de calcul

Raison #2

- Efficacité de certains algorithmes diminue avec le nombre de dimensions.
- Prend en charge la multicolinéarité.

Raison #3

- Aide à visualiser plus facilement les données
- Passage de n dimensions à 2 ou 3

3. RÉSUMÉ DES TECHNIQUES EXISTANTES

- Ratio de valeur manquante (Missing Value Ratio): Si l'ensemble de données comporte trop de valeurs manquantes, nous utilisons cette approche pour réduire le nombre de variables. Nous pouvons supprimer les variables comportant un grand nombre de valeurs manquantes
- Filtre à faible variance (Low Variance filter): Nous appliquons cette approche pour identifier et supprimer les variables constantes de l'ensemble de données. La variable cible n'est pas indument affectée par les variables à faible variance, et donc ces variables peuvent être supprimées sans risque
- Filtre à haute corrélation (High Correlation filter): Une paire de variables ayant une corrélation élevée augmente la multicolinéarité dans l'ensemble de données. Nous pouvons donc utiliser cette technique pour trouver des caractéristiques fortement corrélées et les supprimer en conséquence

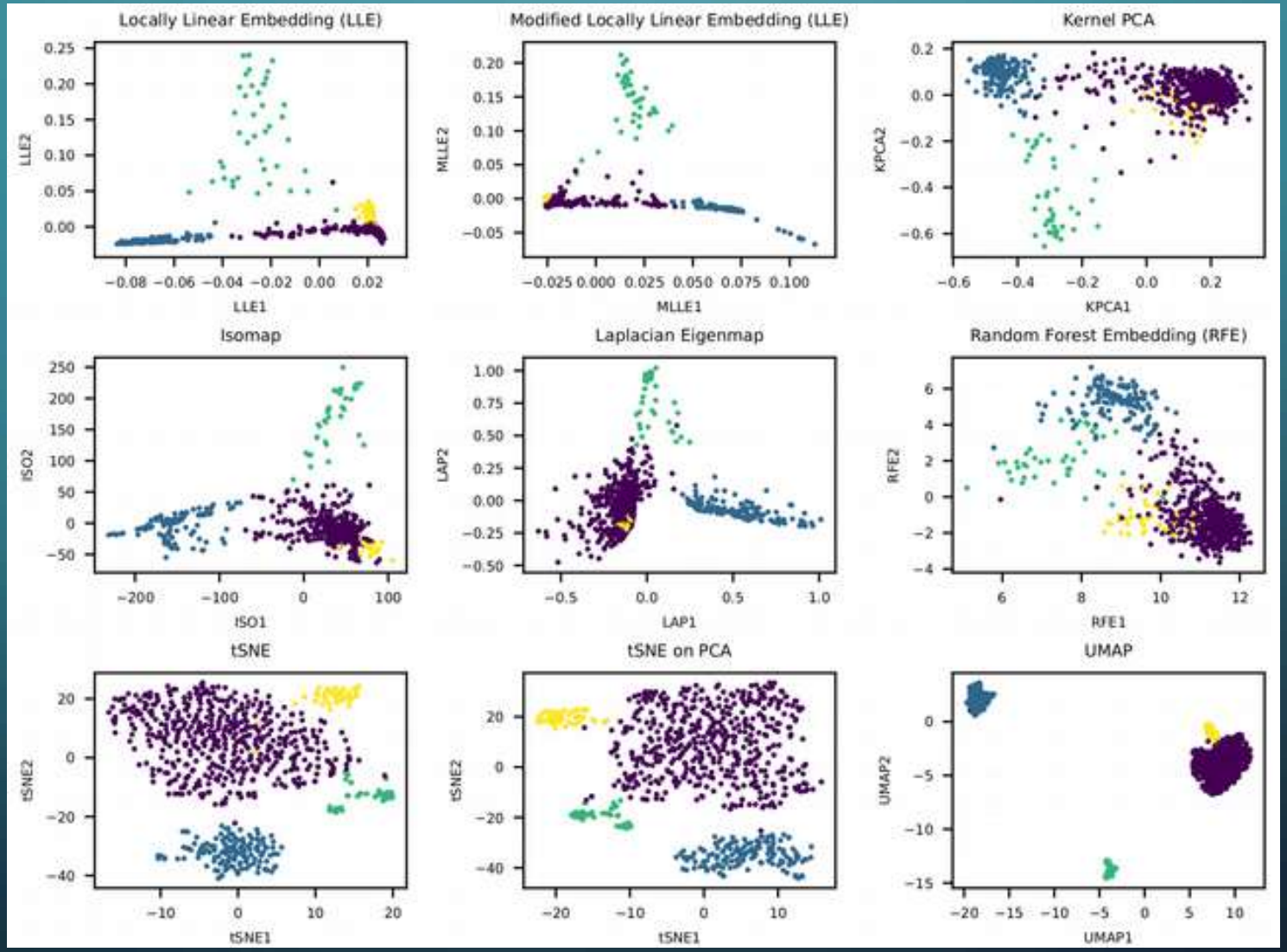
3. RÉSUMÉ DES TECHNIQUES EXISTANTES

- Random Forest : C'est l'une des techniques les plus utilisées qui nous indique l'importance de chaque élément présent dans le jeu de données. Nous pouvons déterminer l'importance de chaque variable et conserver les variables les plus importantes, ce qui entraîne une réduction de la dimension.
- Elimination de variables non significatives en amont (**Backward Feature Elimination**) et de sélection de variables significatives en aval (**Forward Feature Selection**) prennent beaucoup de temps de calcul et sont donc généralement utilisées pour des ensembles de données plus petits.
- Analyse des facteurs (Factor Analysis): Cette technique est la mieux adaptée aux situations où nous avons un ensemble de variables fortement corrélées. Elle divise les variables en fonction de leur corrélation en différents groupes, et représente chaque groupe avec un facteur.
- Analyse en composantes principales (PCA): C'est l'une des techniques les plus utilisées pour traiter les données linéaires. Elle divise les données en un ensemble de composantes qui tentent d'expliquer autant de variance que possible

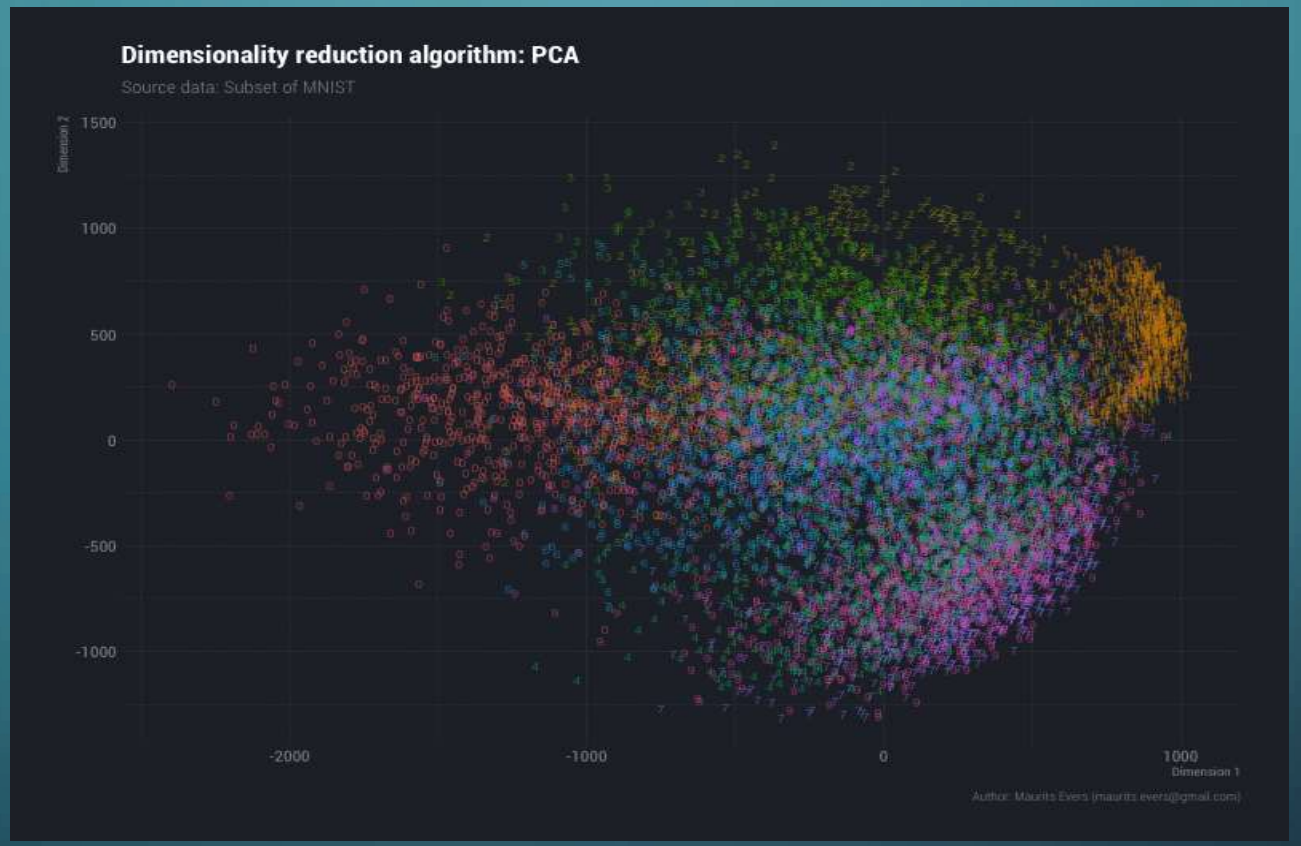
3. RÉSUMÉ DES TECHNIQUES EXISTANTES

- Analyse en composantes indépendantes (ICA): Nous pouvons utiliser l'ICA pour transformer les données en composantes indépendantes qui décrivent les données en utilisant un nombre réduit de variables.
- ISOMAP : Nous utilisons cette technique lorsque les données sont fortement non linéaires
- t-SNE : cette technique fonctionne également bien lorsque les données sont fortement non linéaires. Elle fonctionne aussi très bien pour les visualisations
- LargeVis : une technique qui construit d'abord un graphique du voisin le plus proche K approximé à partir des données, puis dispose le graphique dans l'espace à faible dimension. Par rapport à t-SNE, LargeVis réduit considérablement le coût de calcul de l'étape de construction du graphique et utilise un modèle probabiliste pour l'étape de visualisation, dont la fonction d'objectif peut être optimisé efficacement par une descente de gradient stochastique asynchrone avec une complexité temporelle linéaire.

3. RÉSUMÉ DES TECHNIQUES EXISTANTES



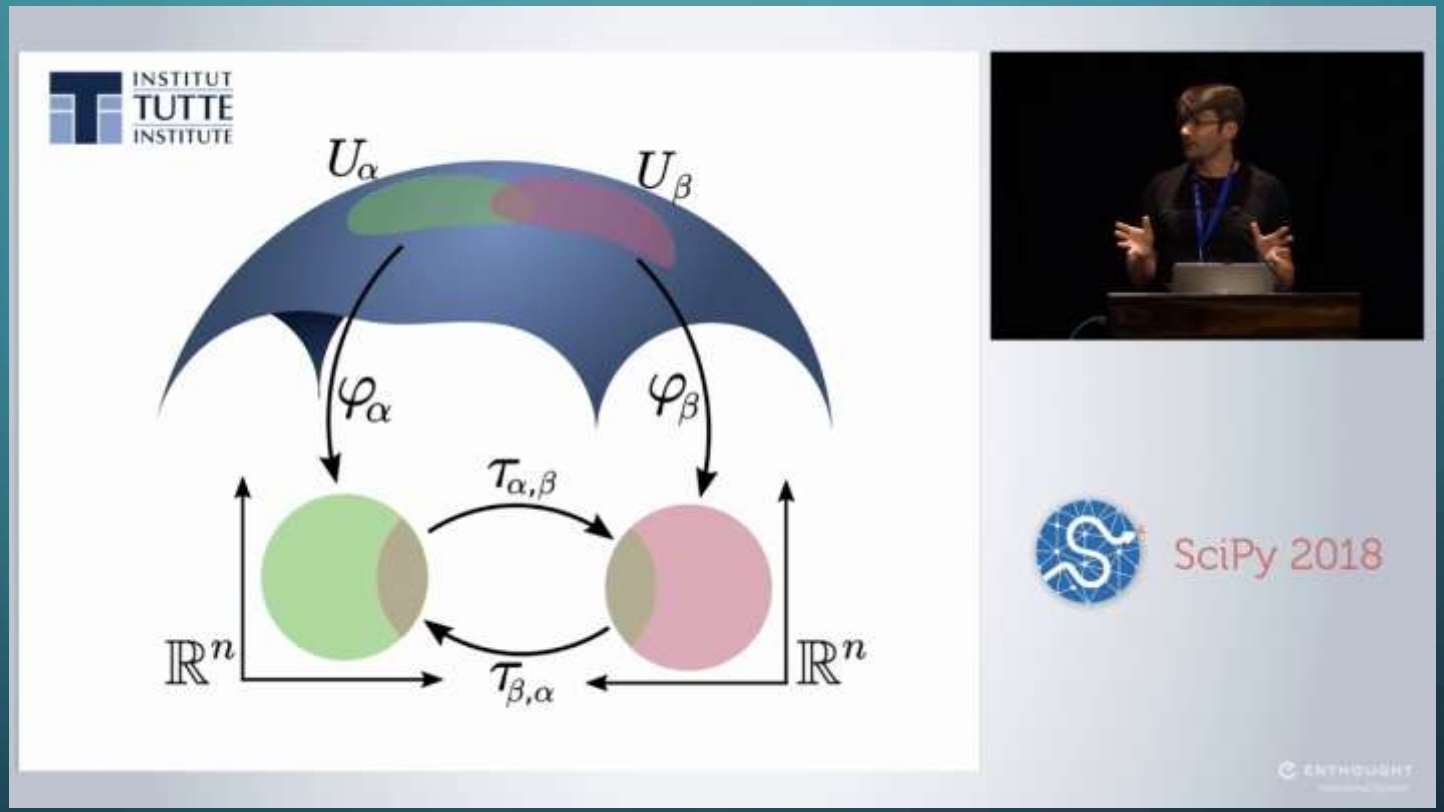
3. RÉSUMÉ DES TECHNIQUES EXISTANTES



4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- La plupart des algorithmes de réduction de la dimensionnalité entrent dans l'une des deux grandes catégories : La factorisation de matrice (telle que l'ACP) ou la mise en forme de graphes (telle que t-SNE). L'UMAP est un algorithme d'agencement de graphes, très similaire à t-SNE, mais avec un certain nombre de fondements théoriques clés qui donnent à l'algorithme une base plus solide.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP



4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

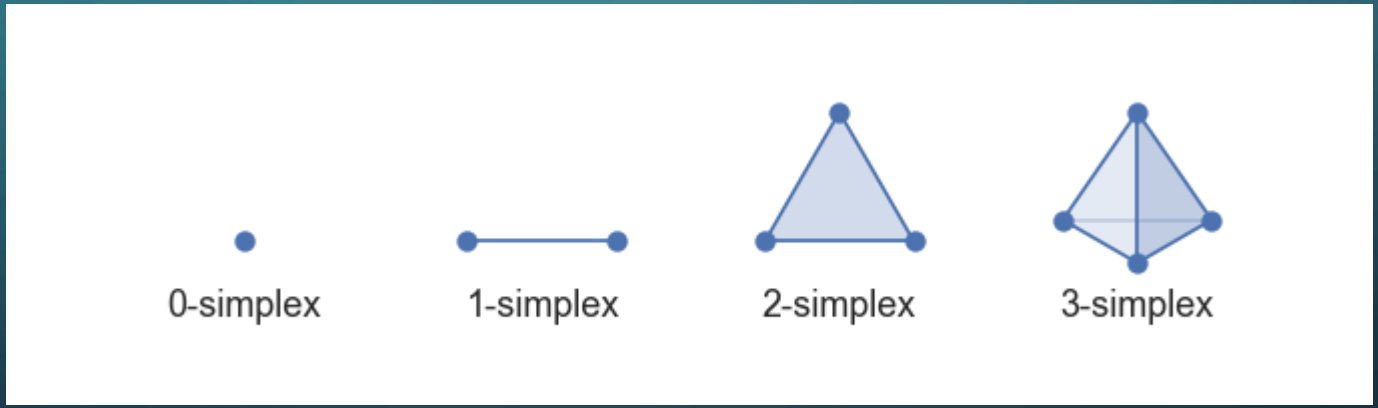
- l'algorithme UMAP se compose de deux étapes :
  - La construction d'un graphe en hautes dimensions
  - Optimisation pour trouver le graphe le plus similaire en basses dimensions.
- Pour atteindre cet objectif, l'algorithme s'appuie sur un certain nombre de connaissances issues de la topologie algébrique et de la géométrie riemannienne.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- Malgré les mathématiques intimidantes, les intuitions derrière les principes de base sont en fait assez simples : UMAP construit essentiellement un graphique pondéré à partir des données de haute dimension, la force des « arêtes » (bords) représentant la "proximité" d'un point donné par rapport à un autre, puis projette ce graphique vers une dimension inférieure.

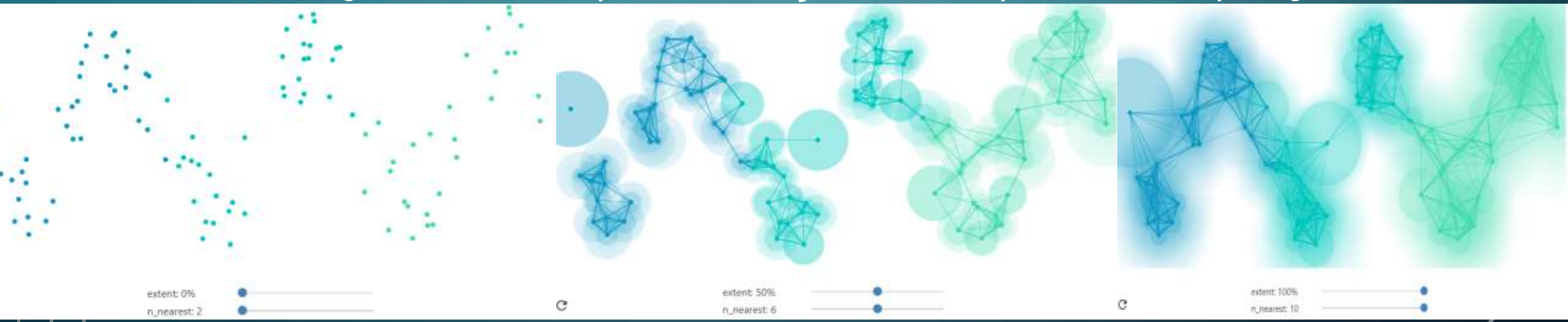
4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- Pour parvenir à la réduction de dimension, l'algorithme utilise un élément de base appelé un simplex. Sur le plan géométrique, un simplex est un objet à k dimensions formé par la connexion de k + 1 points - par exemple, un simplex 0 est un point, un simplex 1 est une ligne et un simplex 2 est un triangle



4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- En considérant nos données comme un ensemble de simplexs, nous pouvons obtenir une représentation de la topologie, et en combinant ces « simplexs » d'une manière spécifique pour former un complexe Čech, nous obtenons certaines garanties théoriques sur la façon dont il représente la topologie.

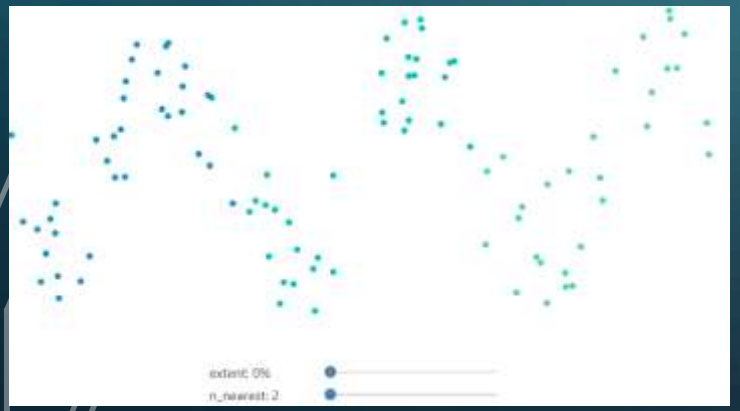


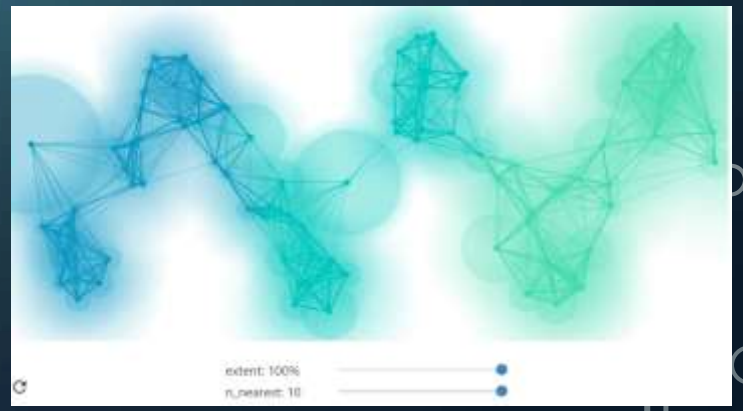
4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- Nous commençons par considérer chaque point de nos données comme un échantillon d'une forme continue et de grande dimension (notre topologie). Nous pouvons considérer chaque point comme un 0-simplex. En étendant à partir de chaque point un rayon r, et en reliant les points qui se chevauchent, nous pouvons construire des ensembles de simplexs à une, deux ou trois dimensions.
- Ce complexe simpliciel fait un travail raisonnable d'approximation de la topologie fondamentale de l'ensemble de données, expliquée par le Théorème des nerfs (Nerve Theorem https://fr.wikipedia.org/wiki/Nerf_d%27un_recouvrement). Il s'avère que la majeure partie du travail de représentation de la topologie est en fait effectuée par les simplexs 0 et 1, qui constituent ce que l'on appelle un complexe de Vietoris-Rips.
- Plus important encore, le complexe de Vietoris-Rips est beaucoup plus facile à calculer, surtout pour les grands ensembles de données. En ne considérant que les simplexs 0 et 1, nous avons en fait construit un graphique qui peut être facilement projeté dans un analogue de plus petite dimension.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- Malheureusement, les données en haute dimension du monde réel présentent un problème que l'UMAP doit surmonter - le choix d'un rayon de la bonne taille.
- Un rayon trop petit et nous tendons vers des groupes de points isolés et locaux. Trop grand, et tout devient connecté. Ce problème est exacerbé par le fléau de la dimensionnalité, où les distances entre les points deviennent de plus en plus similaires dans les dimensions supérieures.
- L'UMAP résout ce problème de manière intelligente : Plutôt que d'utiliser un rayon fixe, UMAP utilise un rayon variable déterminé pour chaque point en fonction de la distance à son kième plus proche voisin. À l'intérieur de ce rayon local, la connectivité est alors rendue "floue" en faisant de chaque connexion une probabilité, les autres points ayant moins de chances d'être connectés.
- Comme nous ne voulons pas qu'aucun point soit complètement isolé, une contrainte est ajoutée selon laquelle tous les points doivent être connectés au moins au point le plus proche. Le résultat final de ce processus est un graphique pondéré, avec des pondérations de bord représentant la probabilité que deux points soient "connectés" dans notre « manifold » à haute dimension.





4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

- La notion locale de distance de chaque point peut être différente de celle de ses voisins, nous devons déterminer si deux points sont connectés en fonction de poids de bords dirigés potentiellement différents. L'UMAP corrige cette incohérence en calculant la probabilité qu'au moins un des bords existe.

- Une fois que le complexe final, complexe simpliciel et flou, est construit, UMAP projette les données dans des dimensions inférieures essentiellement par le biais d'un algorithme « force-directed graph layout ». Cette étape d'optimisation est en fait très similaire à l'expérience t-SNE, mais en sautant à travers les cerceaux théoriques tout en construisant notre complexe simplicial initial, UMAP est capable d'accélérer l'optimisation et de préserver une structure globale plus cohérente que t-SNE.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

AVANTAGES

- Premièrement, il est considérablement plus rapide que t-SNE, où le temps nécessaire à son exécution augmente moins que le carré du nombre de données dans l'ensemble de données. Pour mettre les choses en perspective, un ensemble de données qui pourrait prendre des heures à calculer dans le cas de t-SNE prendra quelques minutes à l'UMAP.

- Le deuxième avantage est que bien que l'UMAP soit également un algorithme stochastique, il est frappant de constater à quel point les projections qui en résultent sont similaires d'une série à l'autre et avec des paramètres différents. Cela est dû, une fois de plus, à l'importance accrue accordée par l'UMAP à la structure globale par rapport à t-SNE. Cela signifie que, contrairement au t-SNE, nous pouvons projeter de nouvelles données sur la représentation à plus faible dimension, ce qui nous permet d'incorporer UMAP dans nos pipelines d'apprentissage machine.

- Le troisième avantage est que l'UMAP préserve à la fois la structure locale et globale. Concrètement, cela signifie que non seulement nous pouvons interpréter deux données proches l'un de l'autre dans les dimensions inférieures comme étant similaires dans les dimensions supérieures, mais nous pouvons également interpréter deux groupes de données proches l'un de l'autre comme étant plus similaires dans les dimensions supérieures.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP

INCONVENIENTS

1. Les hyperparamètres sont très importants

Le choix de bonnes valeurs n'est pas facile et dépend à la fois des données et de vos objectifs (par exemple, le degré de précision de la projection). C'est là que la rapidité d'UMAP est un grand avantage - En exécutant UMAP plusieurs fois avec une variété d'hyperparamètres, vous pouvez avoir une meilleure idée de la façon dont la projection est affectée par ses paramètres.

2. La taille des clusters dans un tracé UMAP ne signifie rien

Tout comme dans t-SNE, la taille des groupes les uns par rapport aux autres est essentiellement dénuée de sens. C'est parce que l'UMAP utilise des notions locales de distance pour construire sa représentation graphique à haute dimension.

3. Les distances entre les clusters peuvent ne rien signifier

De même, les distances entre les grappes risquent d'être non significatives. S'il est vrai que les positions globales des clusters sont mieux préservées dans l'UMAP, les distances entre eux ne sont pas significatives. Encore une fois, cela est dû à l'utilisation des distances locales lors de la construction du graphique.

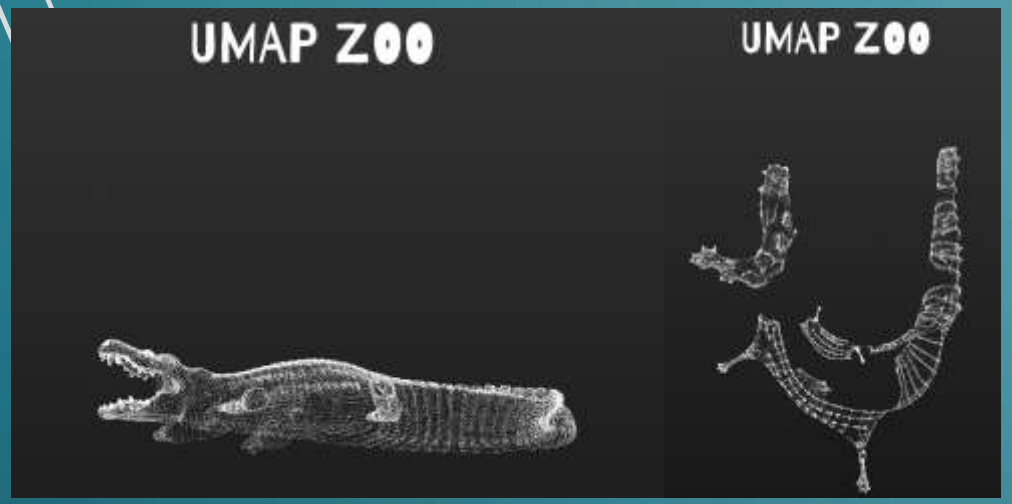
4. Le bruit aléatoire n'a pas toujours l'air aléatoire.

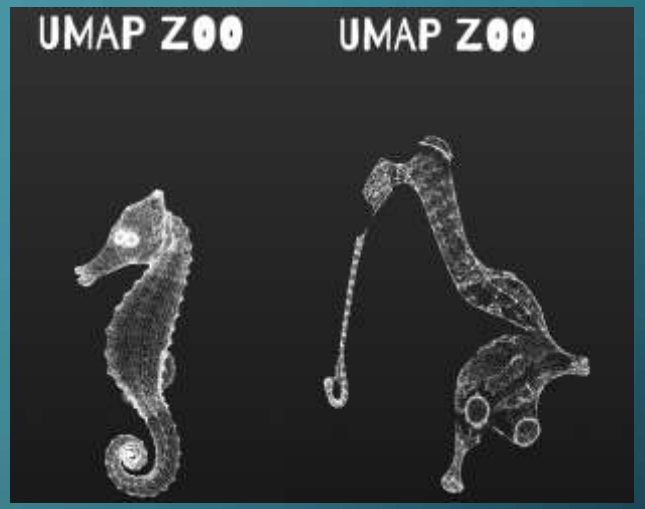
En particulier aux faibles valeurs de n_neighbors, on peut observer des regroupements parasites.

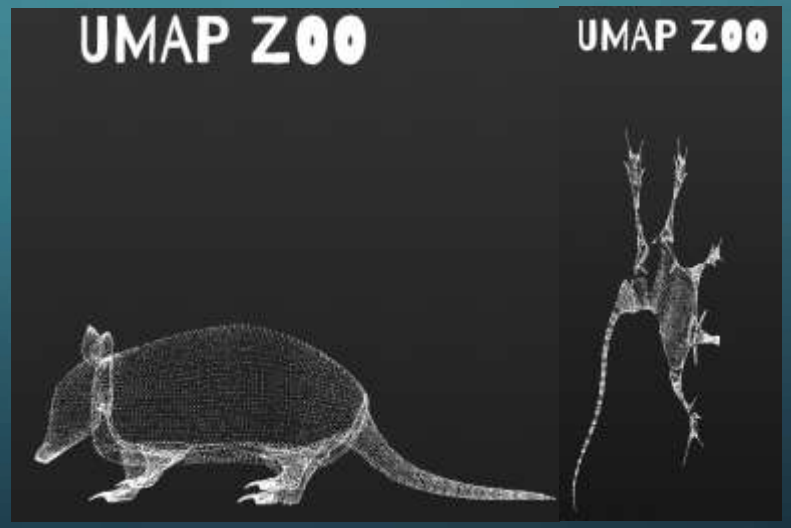
5. Vous pouvez avoir besoin de plus d'un graphique

Comme l'algorithme UMAP est stochastique, différentes exécutions avec les mêmes hyperparamètres peuvent donner des résultats différents. De plus, le choix des hyperparamètres étant très important, il peut être très utile d'exécuter la projection plusieurs fois avec différents hyperparamètres.

4. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION : UMAP







4. VISUALIZATION OF VERY LARGE HIGH-DIMENSIONAL DATA SETS AS MINIMUM SPANNING TREES TMAP

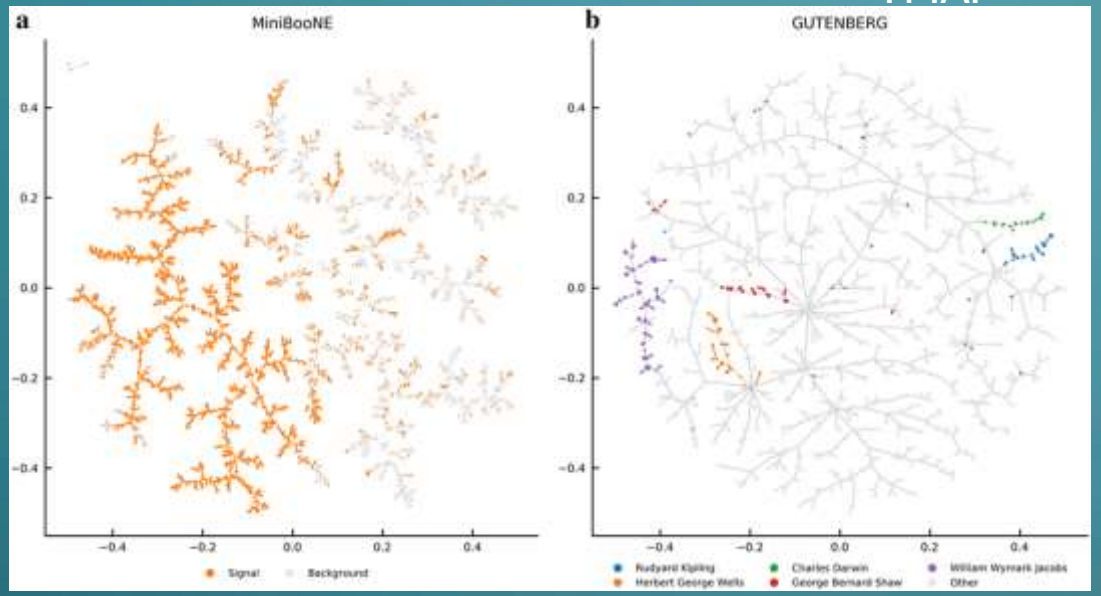
- TMAP est une méthode de visualisation de très grands ensembles de données à haute dimension permettant une grande interprétabilité des données en préservant et en visualisant des caractéristiques à la fois globales et locales.

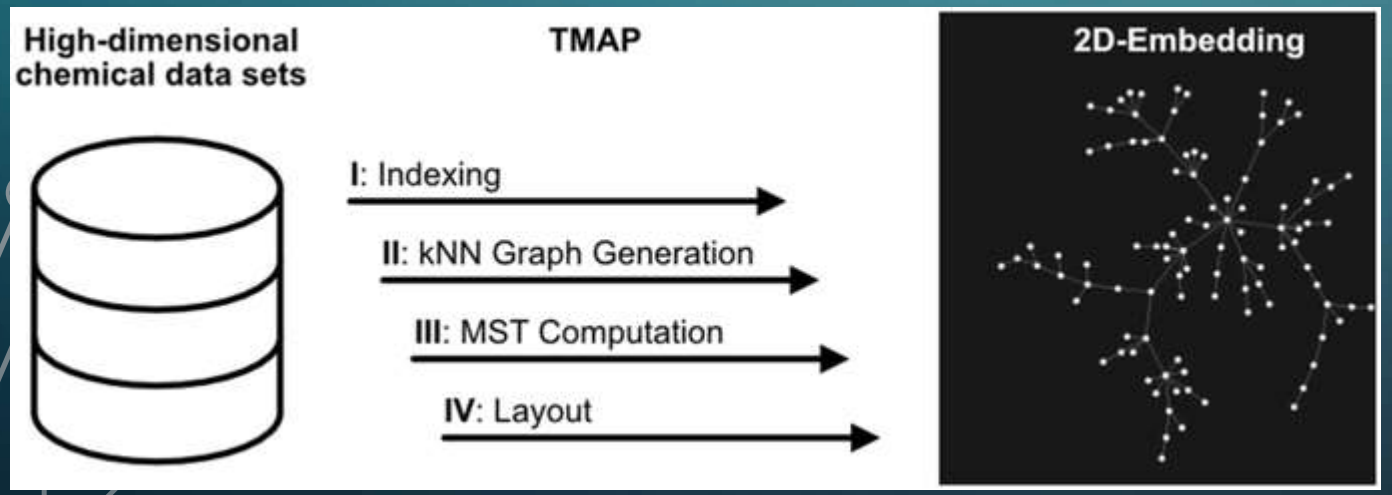
- Visualiser des bases de données de millions de petites molécules organiques et les données de propriétés associées avec un haut degré de résolution, ce qui n'était pas possible avec les méthodes précédentes. TMAP est également bien adapté à la visualisation d'ensembles de données arbitraires tels que des images, du texte ou des données de séquences d'ARN, ce qui laisse entrevoir son utilité dans un large éventail de domaines, notamment la linguistique informatique ou la biologie

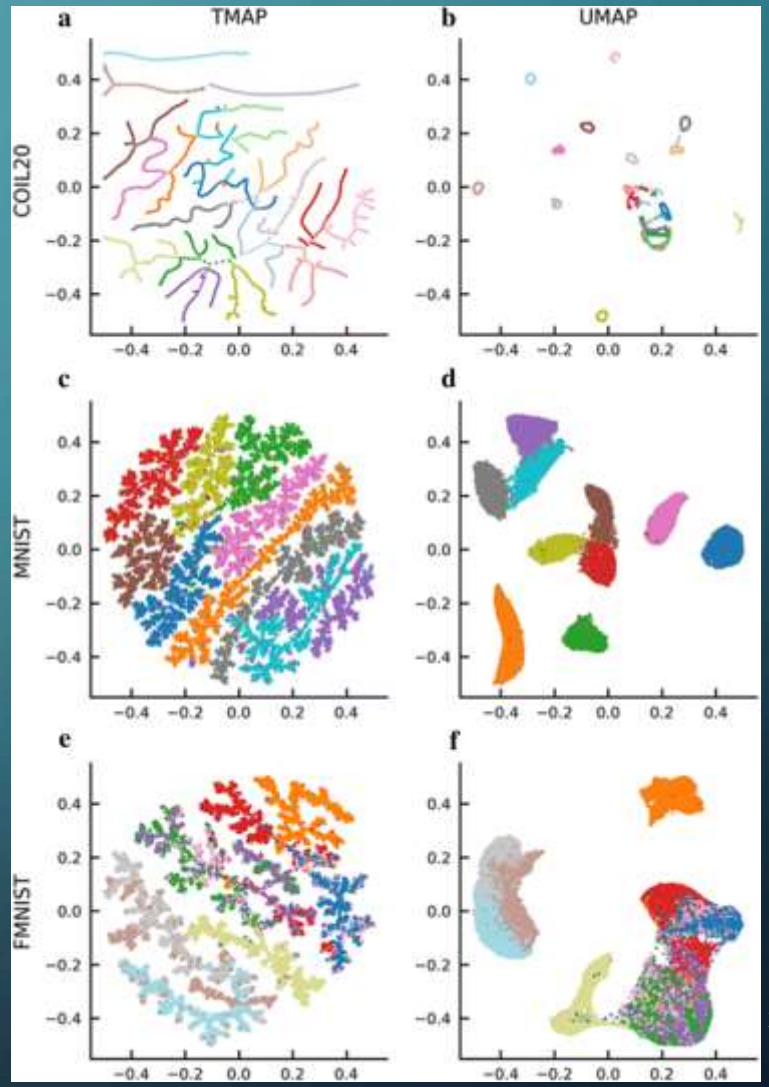
- Le TMAP excelle par sa faible utilisation de la mémoire et sa durée d'exécution, avec des performances supérieures à celles d'autres algorithmes de visualisation tels que le t-SNE, l'UMAP ou le PCA. En ajustant les paramètres disponibles et en tirant parti de la qualité de rendu et de l'utilisation de la mémoire, TMAP ne nécessite pas de matériel spécialisé pour des visualisations de haute qualité d'ensembles de données contenant des millions de points de données.

4. VISUALIZATION OF VERY LARGE HIGH-DIMENSIONAL DATA SETS AS MINIMUM SPANNING TREES

TMAP







MERCI DE VOTRE ATTENTION