PALO IT
Innovation & Transformation

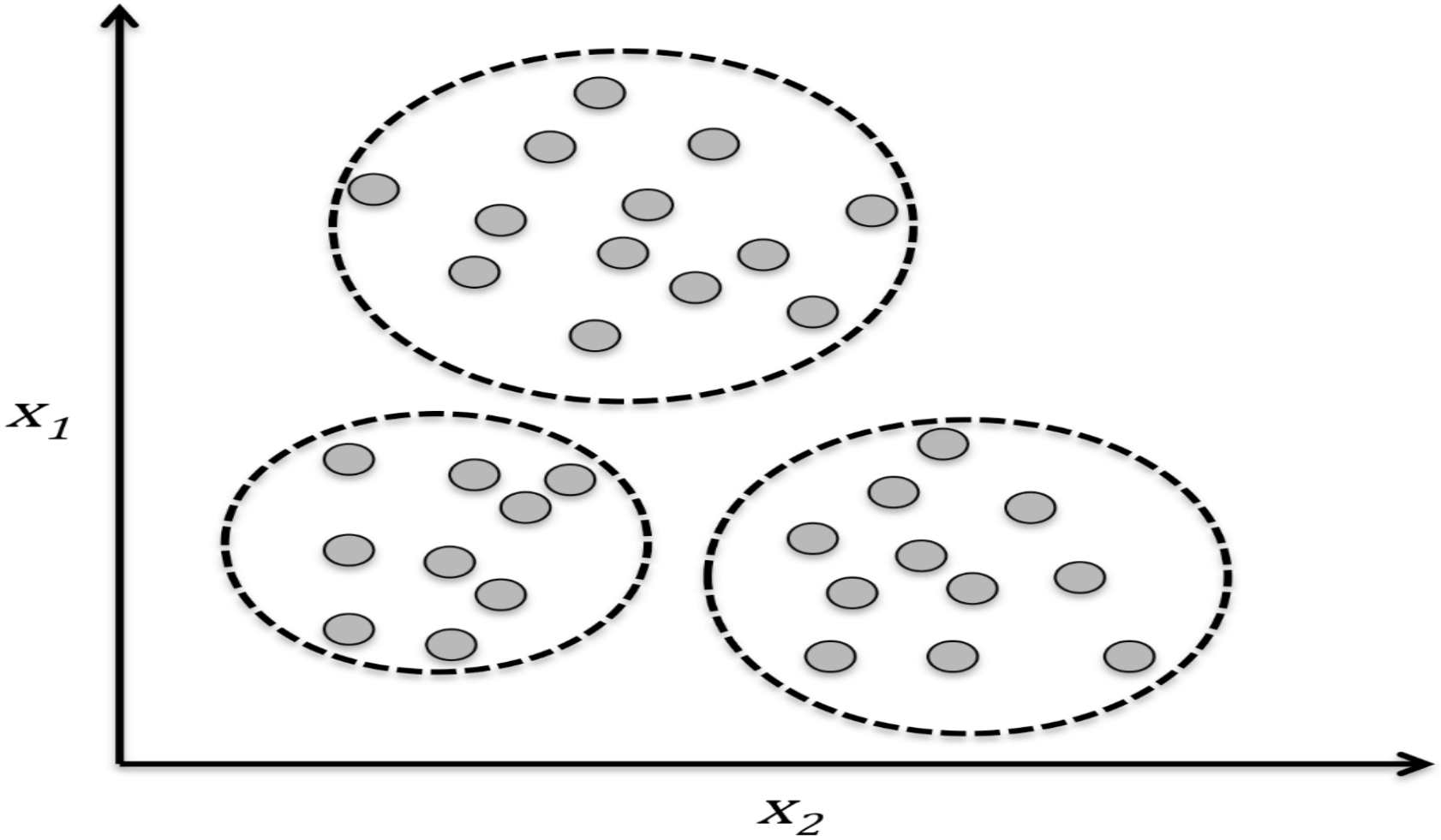
France • Hong Kong • Singapore • Mexico

# Comment faire du non supervisé Avec du supervisé

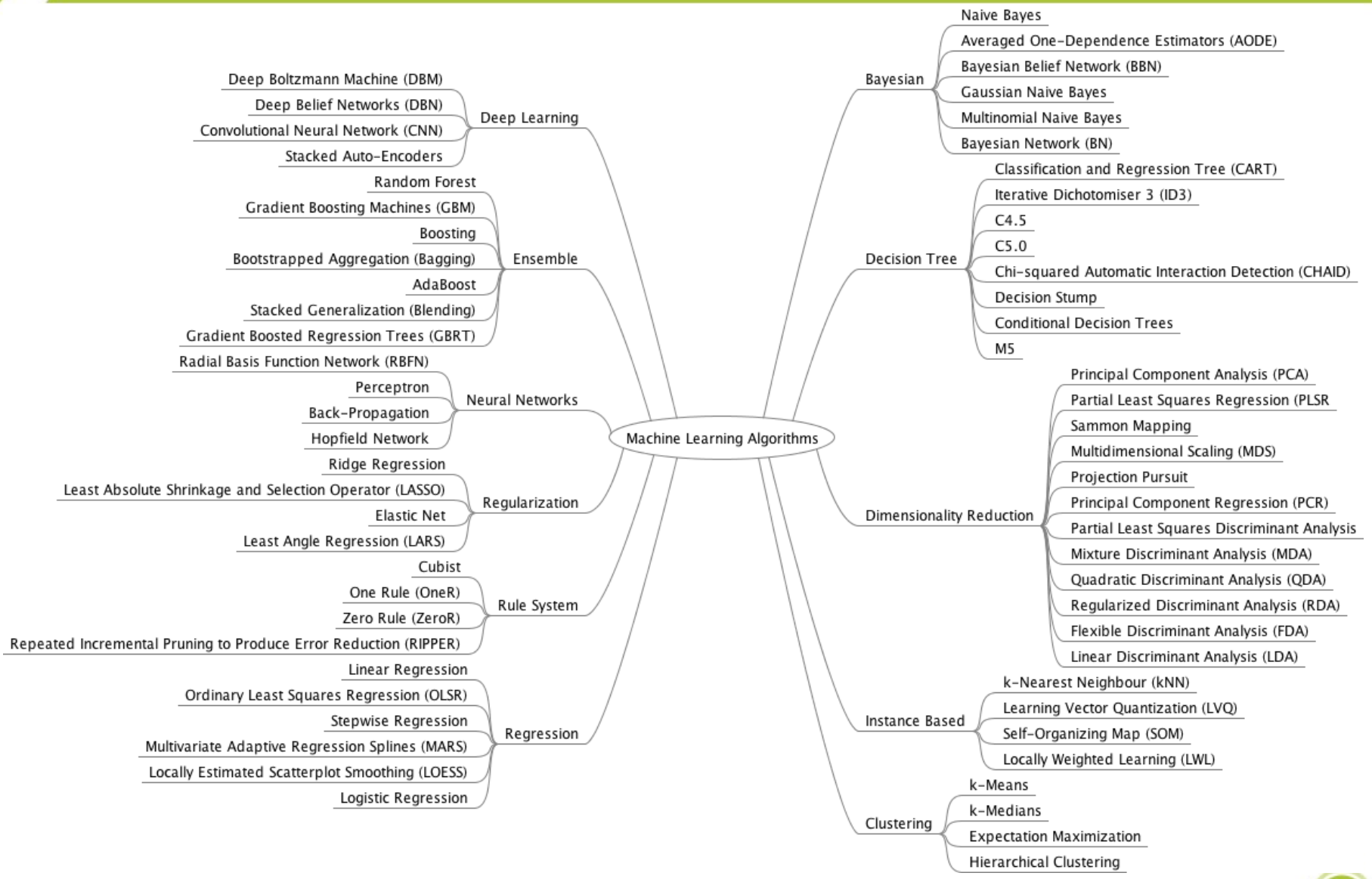
20/01/2017


**Contact**
**Frédéric BERNAROYAT**
Managing Director
01.43.12.63.33 /
06.82.18.29.62
fbernaroyat@palo-it.com

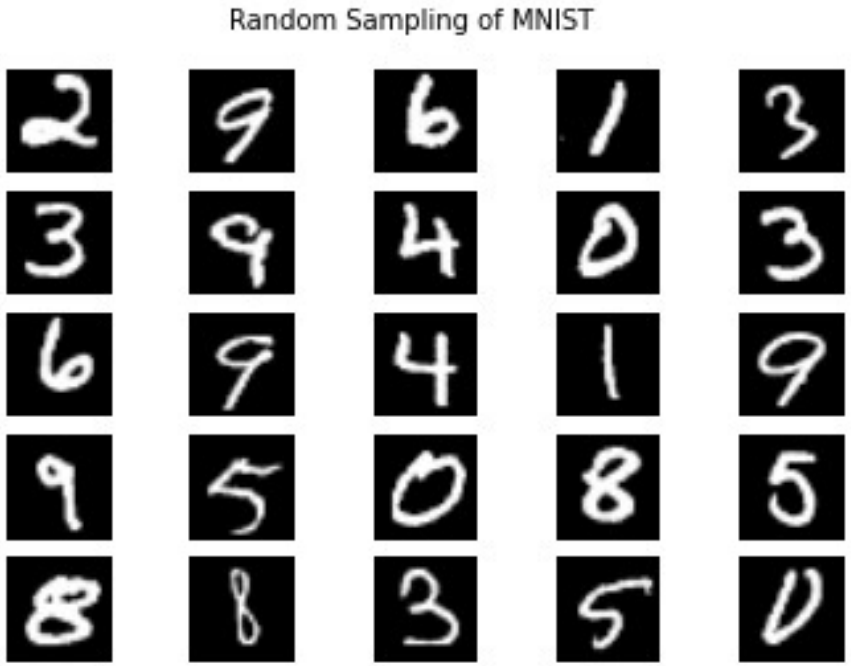
# Non supervisé :Découvrir une classification en groupes non connus



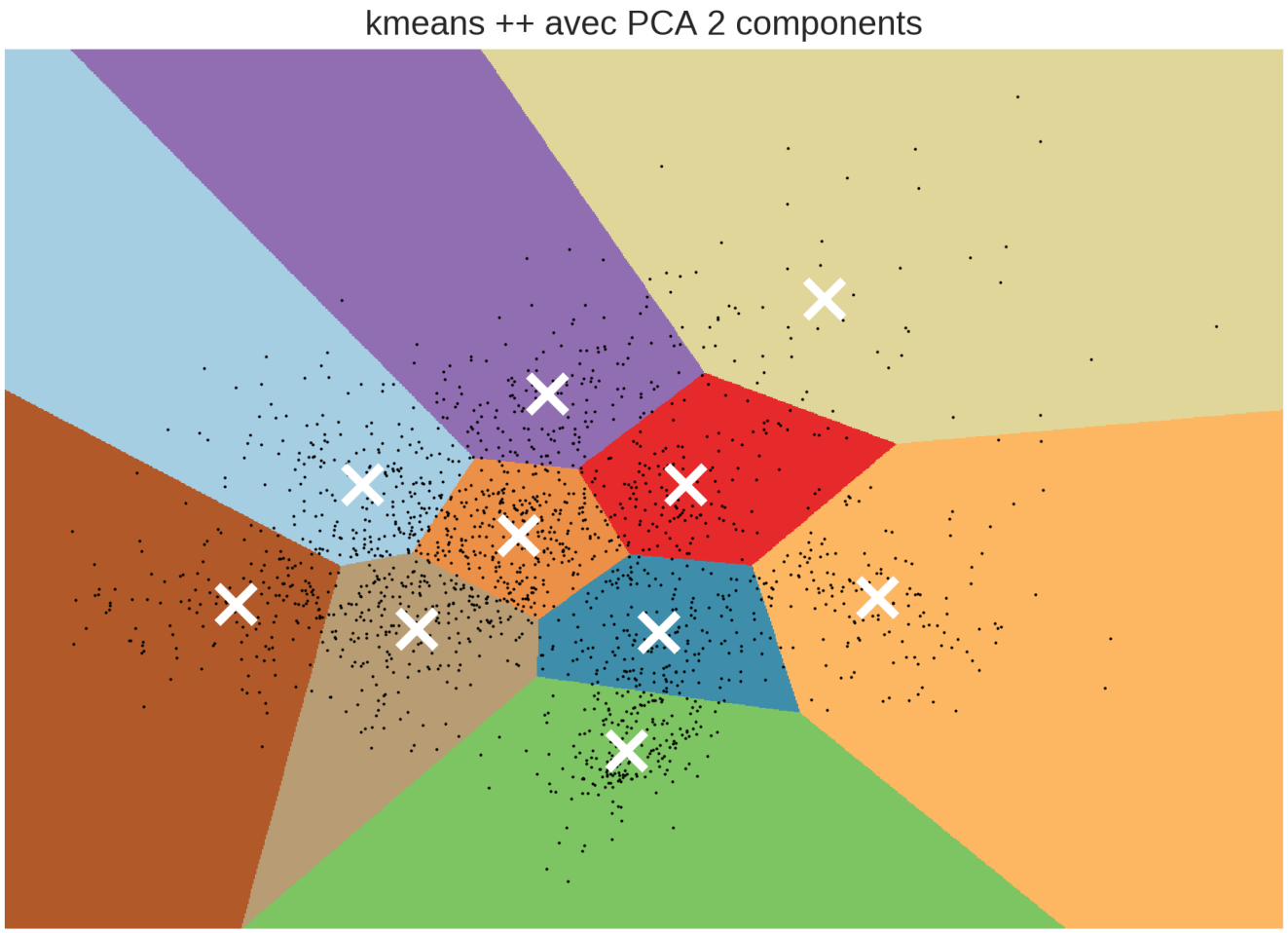
# Vue d'ensemble



# MNIST : 10 chiffres écrits à reconnaître en mode non supervisé

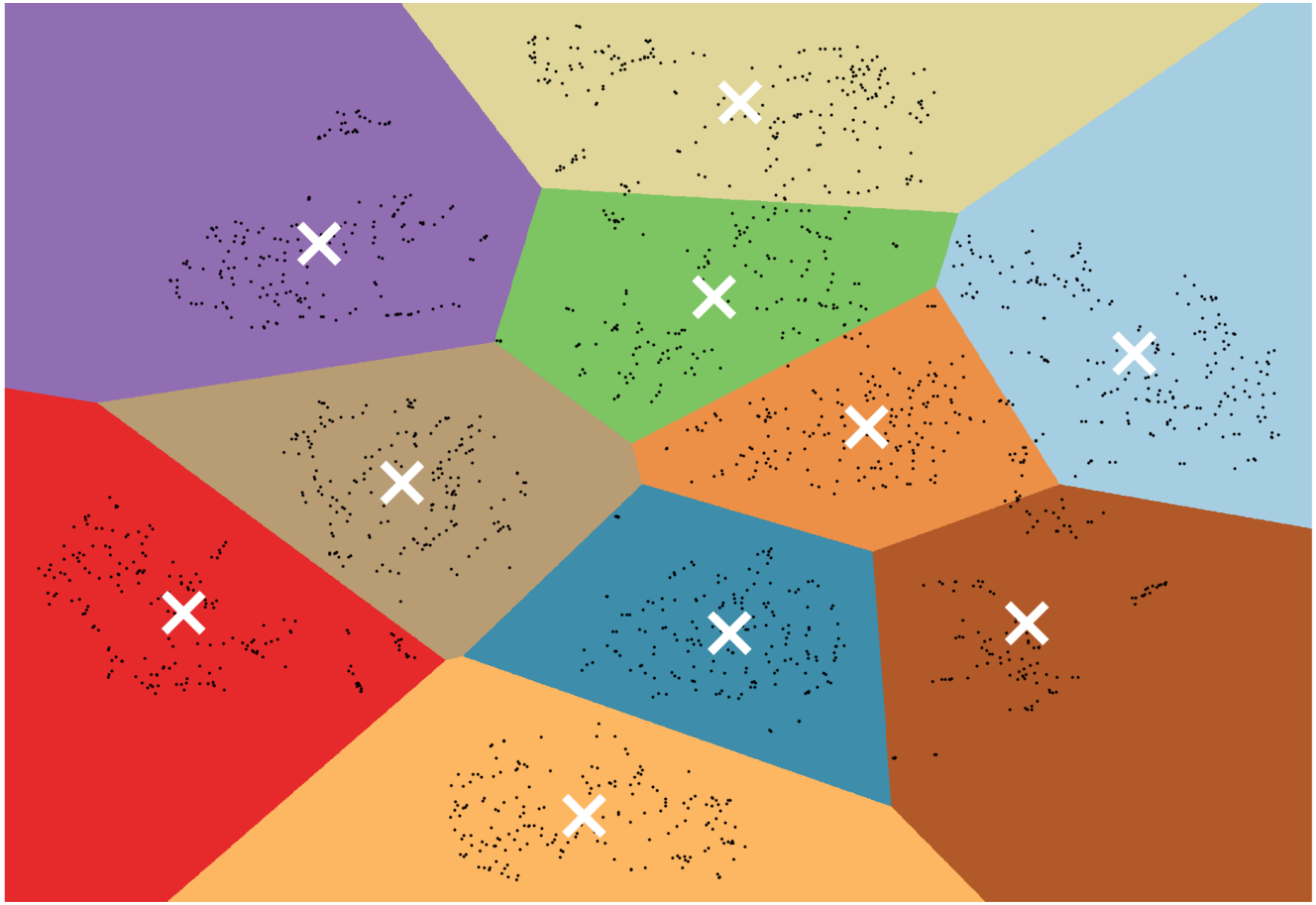


# Résultat standard



# Avec l’aide de RandomForest

With RandomForest analysis before clustering by kmeans ++

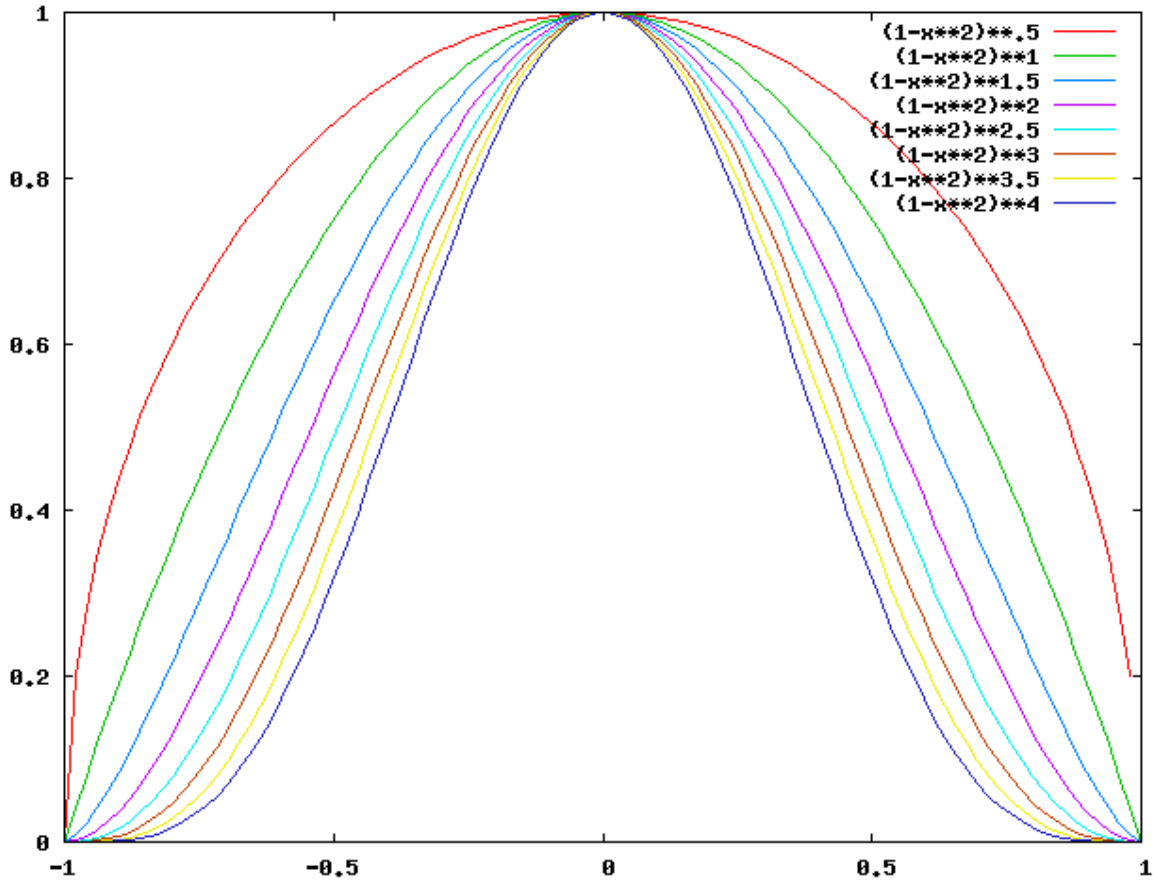


# Comment comparer des  hiboux et des cailloux sans choisir une distance appropriée.


Nous avons l’habitude de prendre les données de toute provenance
Et nous comparons ces mesures en utilisant une distance « classique »

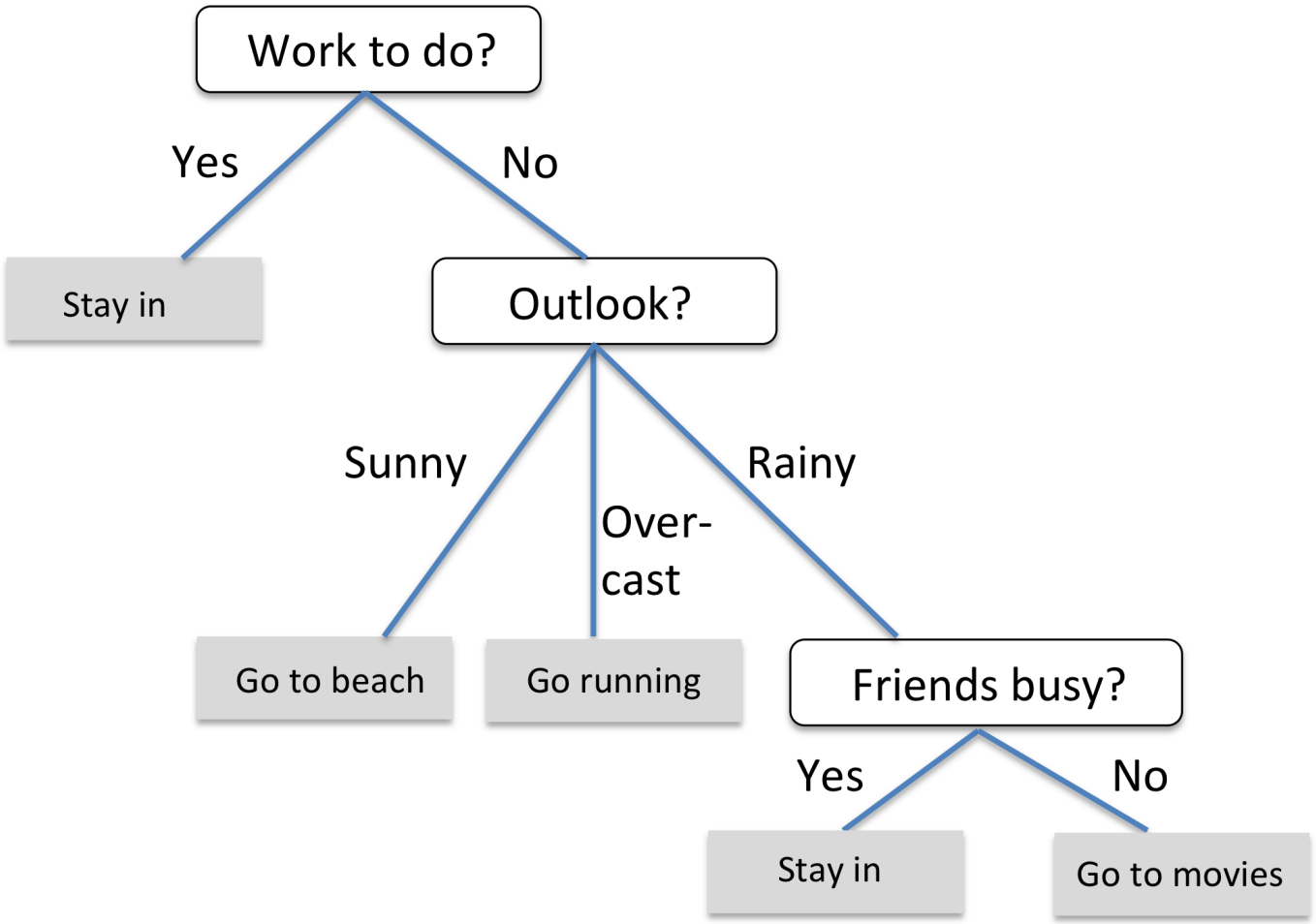
C’est souvent faux…….dans la pratique.

# The curse of dimensionality




Les hyperboules se pincent de plus en plus comme la dimension croît. (Plus précisément, puisque nous intégrons en coordonnées rectangulaires, et que les boîtes rectangulaires circonscrites aux boules s'étendent de plus en plus hors des boules comme la dimension croît, les boules nous paraissent de plus en plus pincées.)

# Arbre de décision : CART



# CART l'idée pour apprendre en supervisé

$$IG(D_p, f) = I(D_p) - \Sigma_j \frac{N_j}{N_p} I(D_j)$$

Ou I(D) est une  fonction qui mesure l impureté d'un groupe D
On utilise

- Gini = Ig
- Entropy = Ih
- Classification error = Ie

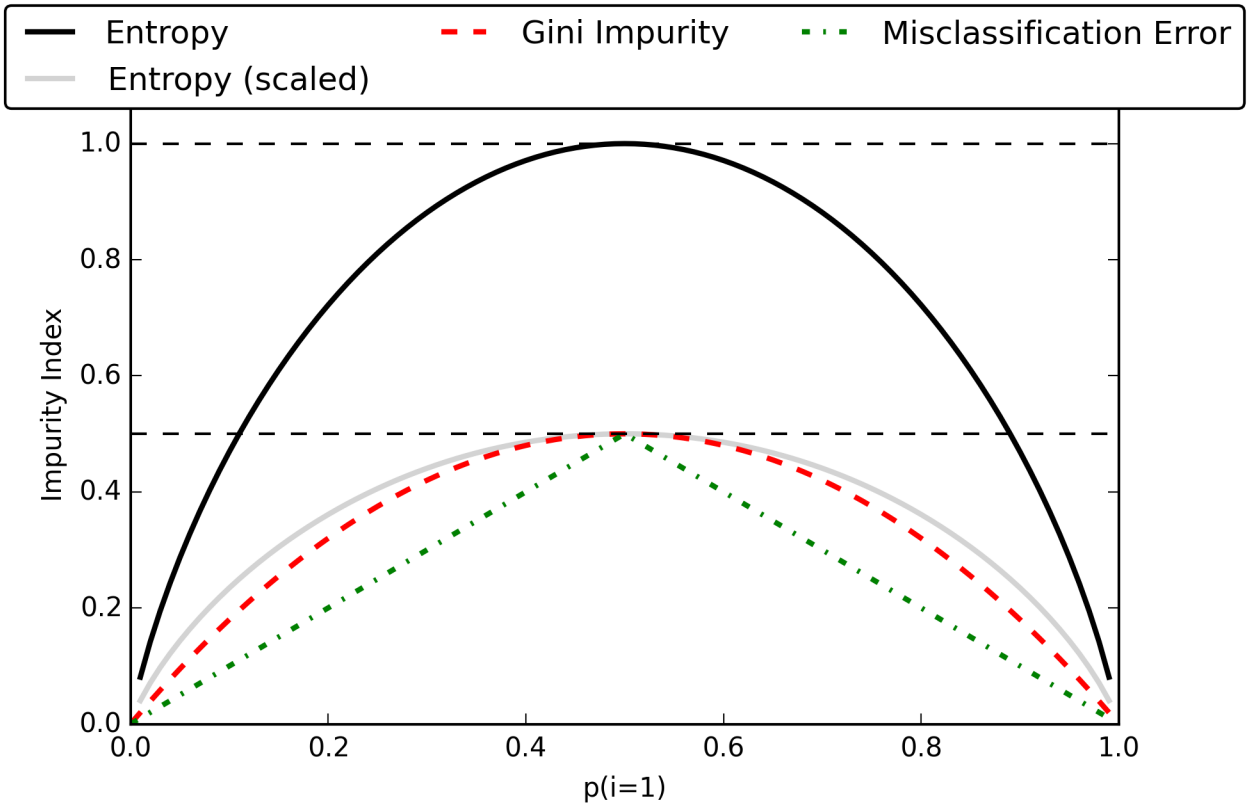
  Ou

$$I_g(t) = \Sigma_i P(i|t)(1 - P(i|t)) = 1 - \Sigma_i P^2(i|t)$$

$$I_h(t) = -\Sigma_i P(i|t) log_2(P(i|t))$$

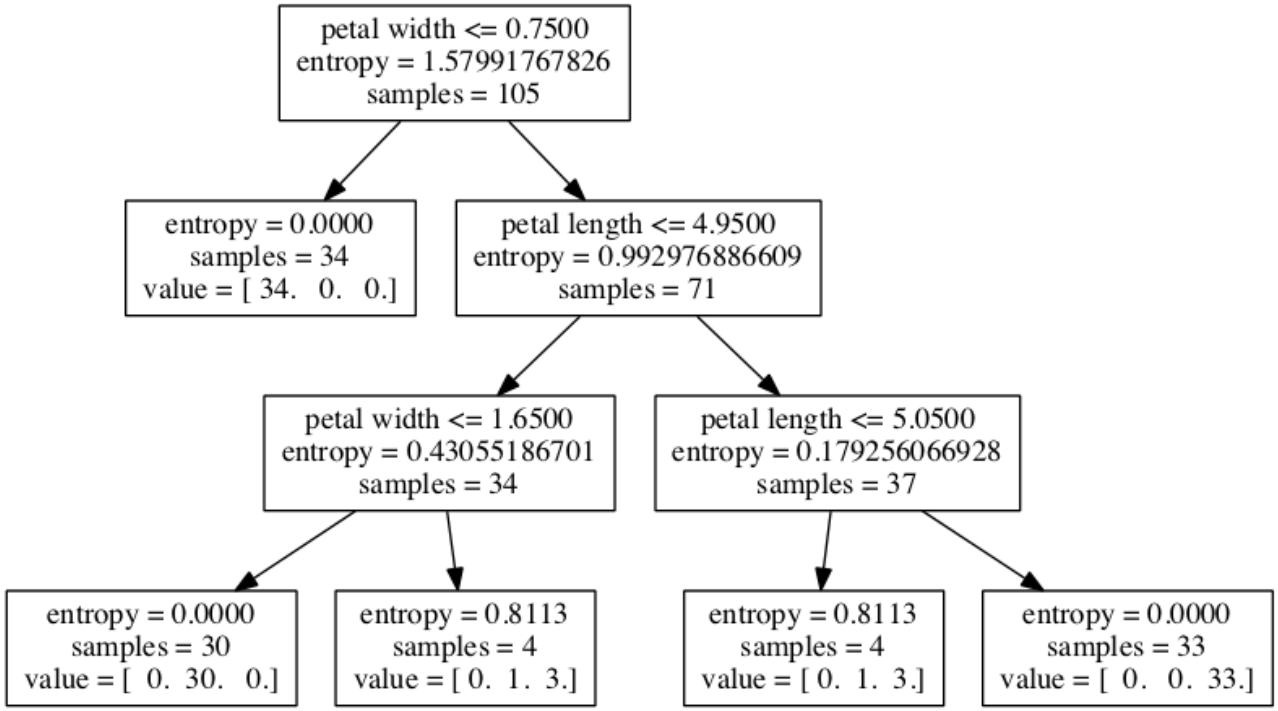
$$I_e(t) = 1 - max_i(P(i|t))$$

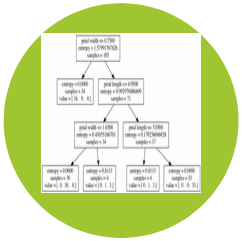
# Avec SKLEARN : CART et ses impuretés




Gini et Entropy sont similaires et maximales si les classes sont parfaitement mélangées

# CART le défaut et la qualité.



**# On sait que le cart « overfitte » si on descend profondément dans l'arbre**
**Mais permet d'expliquer des choix.**

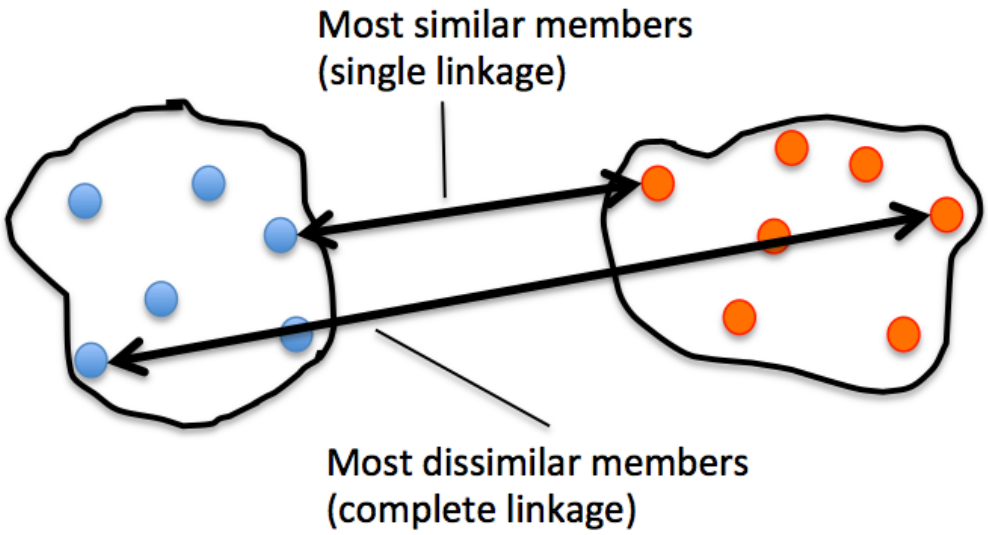
**# A utiliser pour comprendre pas pour prédire**

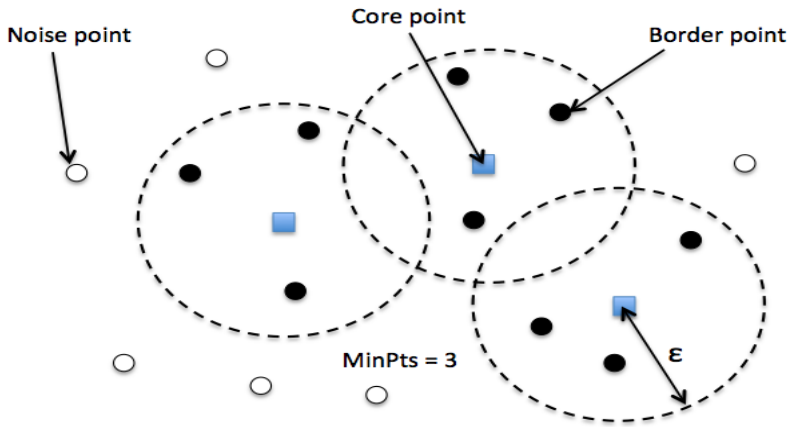
# La forêt d’arbres : une solution qui découvre en surface les patterns.

- **L’idée est de créer N arbre de décisions (CART) en utilisant un sous ensemble de l’échantillon d’apprentissage de taille n basé sur d variables.(d = racine carré de n et n est la taille de l’échantillon, dans SKLEARN).**
  **Agréger le résultat par un vote majoritaire.**

- **Robuste et plutôt facile à utiliser dans les cas non linéaires**
- **De plus cet algorithme permet de comprendre les variables qui expliquent les choix des différents arbres si les résultats sont justes.**

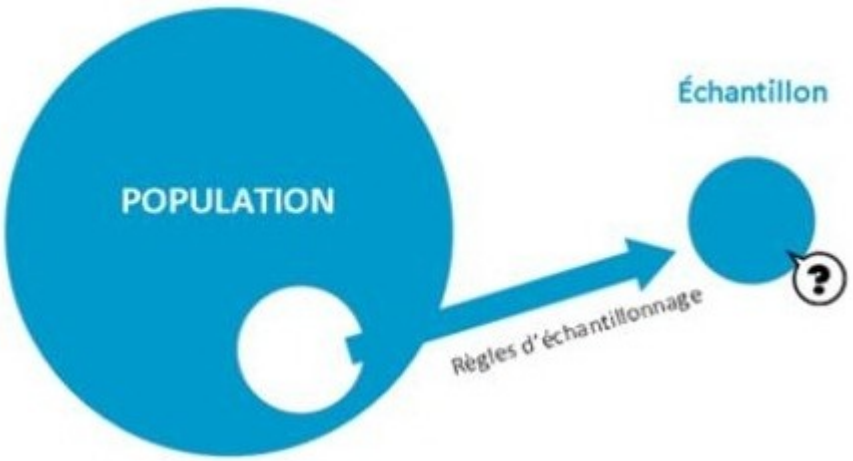
# Classification non supervisée




- Les patterns doivent apparaître en mode non supervisé
- Une idée : utiliser Random Forest qui crée sa propre métrique et n'est pas sensible au nombre de dimensions.

# Création de 2 échantillons : le vrai et le faux.





Groupe A

Groupe B

- **2 modes de création de B : aléatoire ou aléatoire en respectant la distribution de A**
- **Forcer sur le nombre d'arbres mais pas trop.**

# decision_path feature in DecisionTreeClassifier (scikit-learn 0.18.1) fournit une similarité entre individu.

**The decision path of each sample :**
The decision_path  method allows to retrieve the node indicator functions. A non zero element of  indicator matrix at the position (i, j) indicates that the sample i goes  through the node j.

Exemple :
node_indicator = estimator.decision_path(X_test)

Similarly, we can also have the leaves ids reached by each sample.

leave_id = estimator.apply(X_test)

**Update si nécessaire:**
conda update scikit-learn

# Passer d'une similarité binaire à une métrique sur 2 dimensions

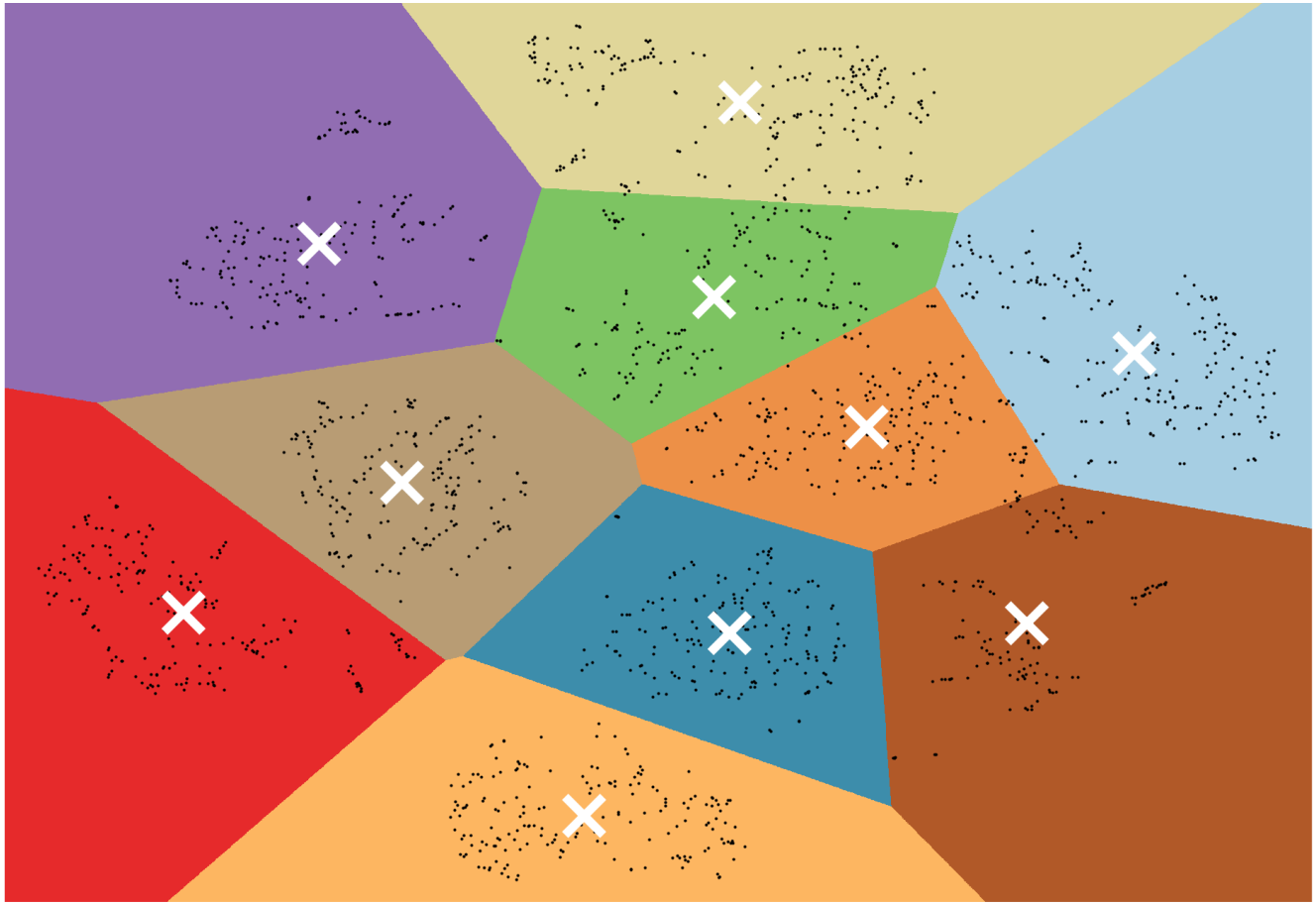
TSNE : t-distributed Stochastic Neighbor Embedding.
Un outil pour visualiser des points relier par des similarités.

( minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding )

$$d(P,Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)})$$

# Faire un « kmeans » sur les données projetées en 2 dimensions :voir  le « notebook »

With RandomForest analysis before clustering by kmeans ++



PALO IT
Innovation & Transformation


**France**
3, rue Scribe
75009 Paris
+33(0)1 43 12 63 30
france@palo-it.com

**Singapore**
56B Boat Quay
Singapore 049845
+65 6220 9908
singapore@palo-it.com

**Hong Kong**
21/F, 151 Hollywood Road
Central, Hong Kong
+ 852 3711 3225
hongkong@palo-it.com

**Mexico**
Calle Moliere 50,
11560 Mexico CDMX
+52(1) 55 4000 1282
mexico@palo-it.com


www.palo-it.com