EPSI l'École d'ingénierie informatique

Machine Learning

Collecte, préparation des données et mise en œuvre

Cours 2

Jean-Claude Houbart

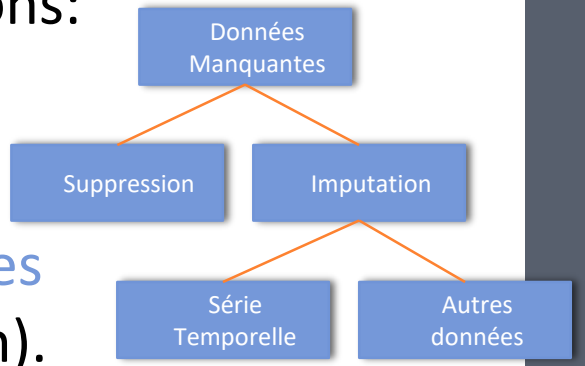
EPSI l'École d'ingénierie informatique

Compléter ses données

Jean-Claude Houbart

Traitement des données manquantes

- Pas de solution idéale.
- Essayer de comprendre pourquoi les données manquent. Corriger si possible.
- Deux stratégies : supprimer ou compléter.
- La suppression est possible à deux conditions:
  - La proportion du volume de données à supprimer est faible.
  - Les données manquantes sont distribuées aléatoirement (MAR, Missing At Random).
- Dans les autres cas, il est préférable de compléter les valeur manquantes (imputation) avec deux cas de figures:
  - Série Temporelle
  - Valeurs quelconques



Suppression de données

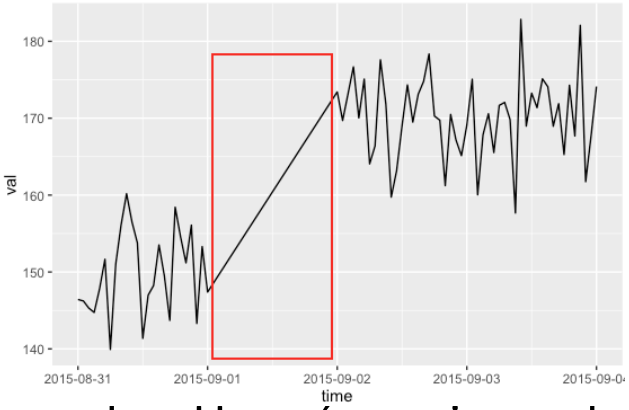
- Cas le plus courant : Suppression des lignes contenant des valeurs manquantes.
- Si il y a plus de 60% de valeur manquante pour une colonne, on peut la supprimer surtout si elle est fortement corrélée à une ou plusieurs autres colonnes:
  - Exemple : CA et Effectif des entreprises.
- La suppression par paires (Pairwise deletion), consiste à supprimer les lignes seulement en fonction de ce qu'on est en train de calculer.
- Par exemple, si on calcule la corrélation entre 2 colonnes, on va uniquement supprimer pour ce calcul les lignes où une des deux valeurs manquent. Il faut donc faire le travail de suppression pour chaque type de calcul.

Imputation de séries temporelles (1)

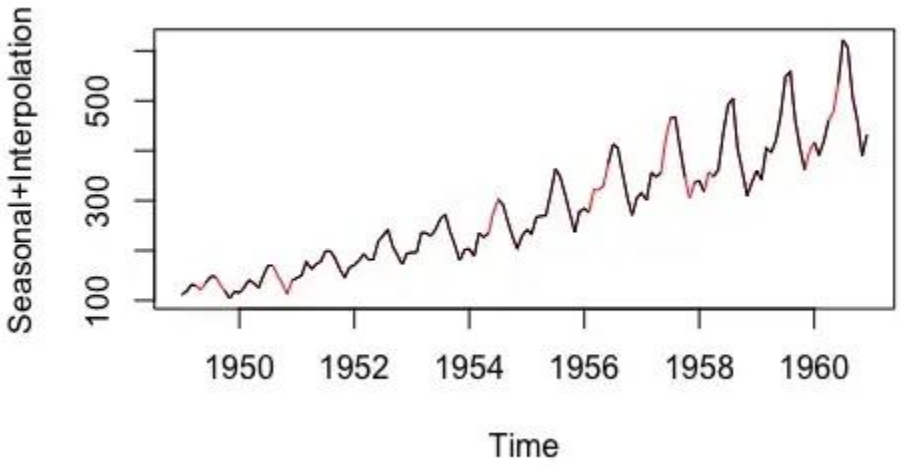
- Il faut tenir compte de la :
  - La tendance : Les données suivent une progression dans le temps.
  - La saisonnalité : Les données ont un comportement différent suivant la période de l'année.
- Si les données n'on ni tendance ni saisonnalité, alors remplacer les valeurs par une grandeur statistique :
  - Moyenne
  - Médiane (autant en dessous qu'au dessus)
  - Mode (valeur la plus fréquente)
  - Aléatoire dans un intervalle bien choisi (par exemple moyenne +-ecart-type).

Imputation de séries temporelles (2)

- S'il y a une tendance, Une interpolation linéaire suffit généralement. Exemple:



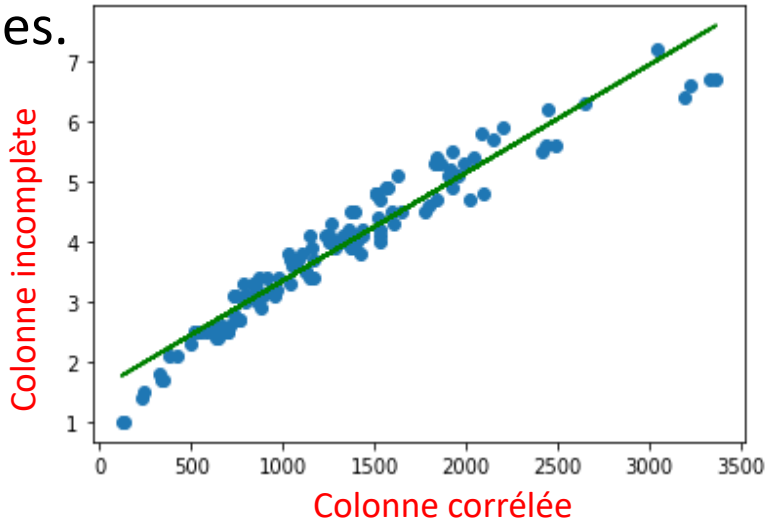
- S'il y a une saisonnalité en plus, Une régression polynomiale est préférable



- Il y a des méthodes plus évoluées dans le package imputeTS

Imputation de données quelconque

- Nous avons encore deux cas : données discrètes (« categorical » : choix dans une liste de valeurs) ou continues
- Pour les données discrètes, on peut créer une nouvelle classe « inconnu ».
- Pour les données continues, on peut utilisé une valeurs statistique : Moyenne, médiane, mode ou aléatoire.
- Dans les deux cas, on peut utiliser une régression pour déduire les données manquantes en partant d'une ou plusieurs autres colonnes bien corrélées.



EPSI l'École d'ingénierie informatique

Créer des Features

Jean-Claude Houbart

Création / Extraction de Features

- Extraction consiste à déduire des nouvelles valeurs de features existant.
- Technique particulièrement adaptés aux textes, mais aussi sur un numéro de téléphone, un code postal…
- Le TP contient un exemple de ce type avec le champ Titre des passagers du Titanic:

PassengerId	Name
87	Slocovski, Mr. Selman Francis
193	Navratil, Master. Michel M
183	Becker, Master. Richard F
816	Heininen, Miss. Wendla Maria
46	Lennon, Mr. Denis
369	Aubart, Mme. Leontine Pauline
596	Leitch, Miss. Jessie Wills
647	Simonius-Blumer, Col. Oberst Alfons
228	Fahlstrom, Mr. Arne Jonas
294	Mineff, Mr. Ivan

Création de Features / Transformation

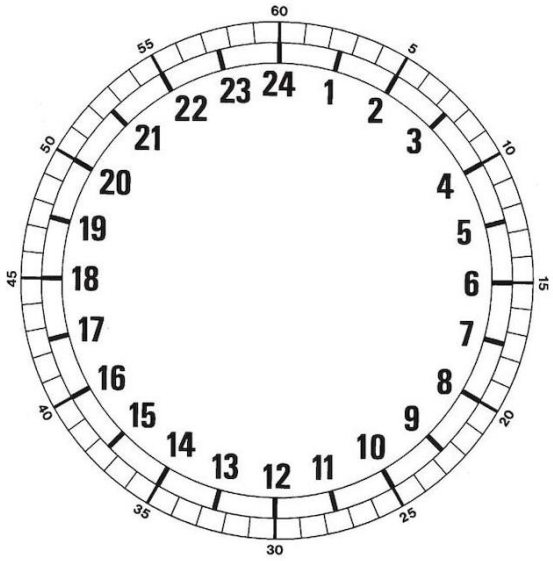
- Une nouvelle feature peut aussi être générée à partir d'un ou plusieurs feature existant.
- Le ratio entre deux valeurs est un classique.
  - Exemple pour un modèle e-Commerce : % panier/page produit.
- Compter des conditions => Exemples
- Group & Transforms permet de créer un nouveau feature issue de regroupement des exemples selon un critère puis d'une grandeur statistique sur ce groupe.
  - Exemple : Pour des statistiques de vente, ajouter à chaque exemple la moyenne des ventes pour la région.

=> Exemples

- Dans le TP un nouveau feature sera construit à partir des features existant sibsp et parch représentant le nombre frères/soeur/époux et de parents/enfants présents à bord pour le passager.

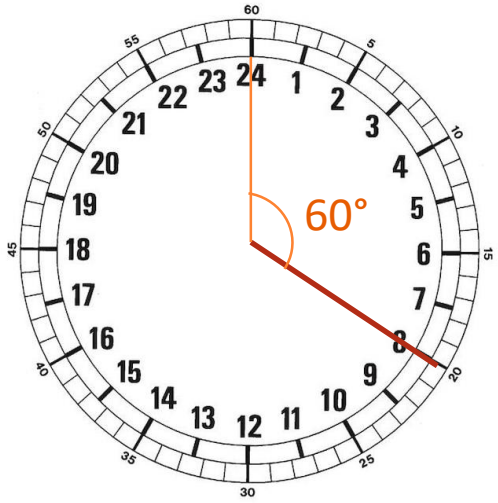
Données Cycliques

- Certaines données sont cycliques. Par exemple l'heure, le numéro de jour dans l'année, une direction…
- On ne peut pas simplement utiliser cette données en la normalisant. En effet, la valeur minimale est en fait proche de la valeur maximale (Exemple : 24 plus proche de 1 que de 20).

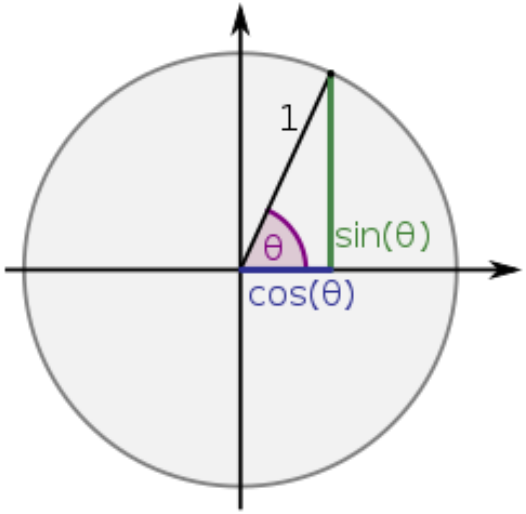


Données Cycliques

- Certaines données sont cycliques. Par exemple l'heure, le numéro de jour dans l'année, la direction…
- On ne peut pas simplement utiliser cette données en la normalisant. En effet, la valeur minimale est en fait proche de la valeur maximale (0° et 359° pour la direction en par exemple).
- Solution : une peu de trigonométrie. 2 variables générées.



x = 8 heures du matin



60° (4h/24)

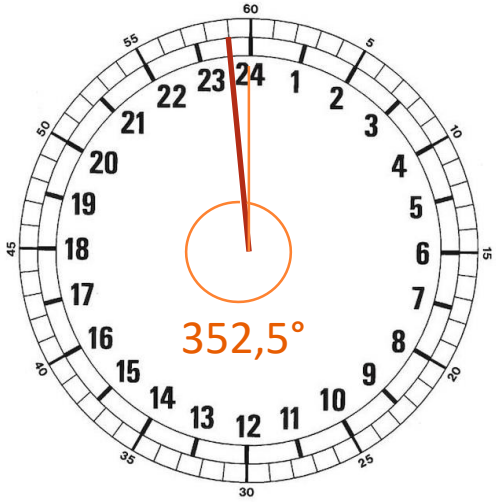
$\frac{\pi}{6}$ radians

x_cos = 0,5

x_sin = 0,866

Données Cycliques : Valeurs proches

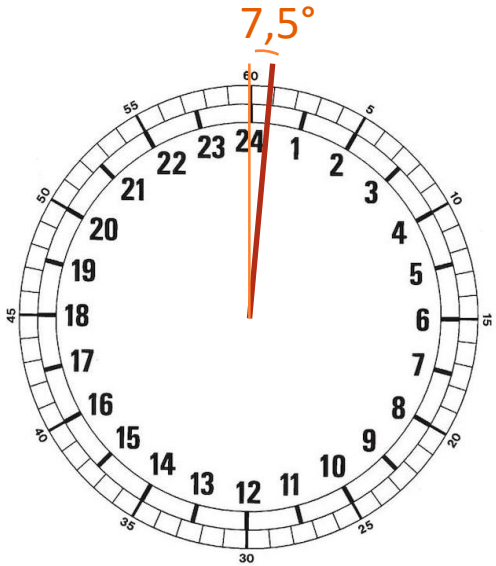
- Deux valeurs proches (23h30 et 0h30) d'une variable cyclique génèrent bien deux couples de valeurs proches.



x=23h30

x_cos = 0,99

x_sin = -0,13



x=0h30

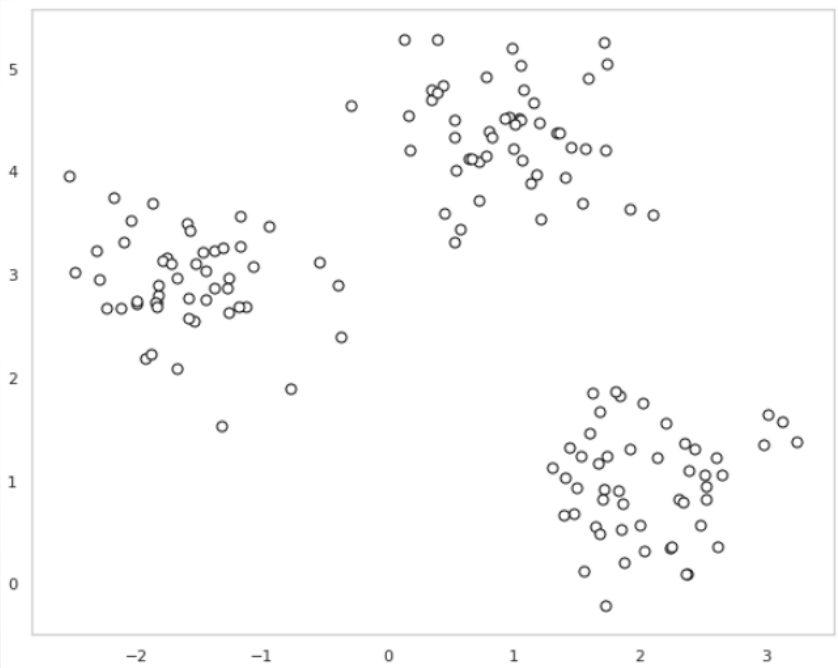
x_cos = 0,99

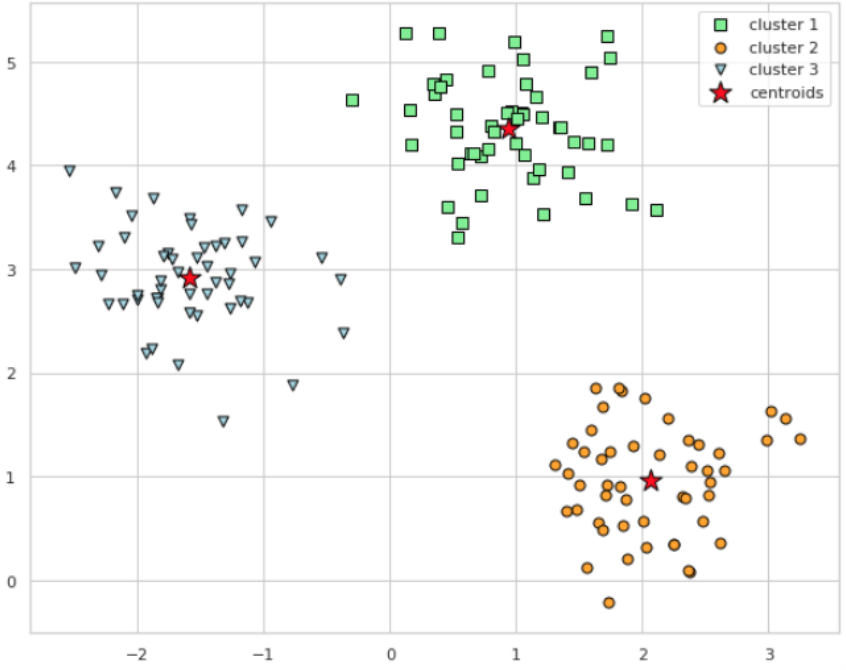
x_sin = 0,13

Clustering

Regroupement les données (en N dimensions) en un nombre quelconque de clusters.

Apprentissage non supervisé.

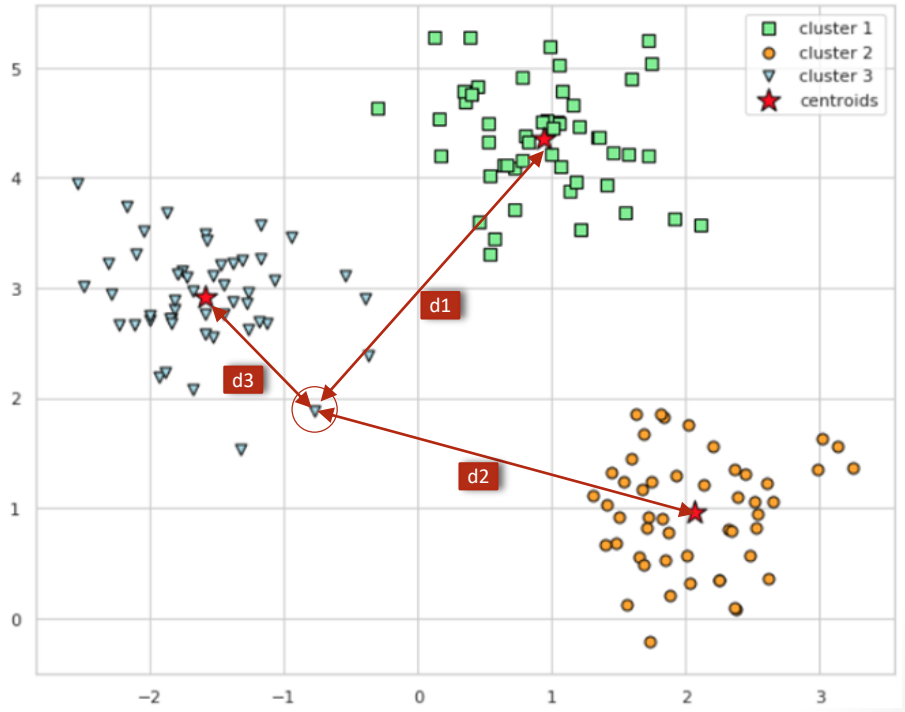




Création de Features / Clustering

- Sur les données d'apprentissage, un algorithme de clustering, typiquement kMeans, regroupe les exemples autour de N centroïds, format ainsi N clusters.
- Un nouveau feature est alors crée constitué du numéro de cluster de chaque exemple. N° cluster = 3 dans l'exemple

- Alternative : On crée N feature, pour chaque distance entre l'exemple et un centroïd.
- Exemple pour le point encerclé :
  - Feature 1 : d1
  - Feature2 : d2
  - Feature 3 : d3



Target Encoding

- Le target encoding utilise une statistique sur le label pour créer un nouveau feature.
- On peut se servir de la combinaison Group & Transform vue précédemment.
- Pour éviter les problèmes sur les petites catégories, il est d'usage de faire un mixte entre la moyenne pour la valeur de feature et la moyenne globale du label :

  Feature = coef * moyenne catégorie + (1-coef) * moyenne globale

- Le coefficient peut être calculé par la formule :

  coef = n / (n + m)

  … où n est le nombre d'éléments et m le facteur de lissage.

  => Exemples

EPSI l'École d'ingénierie informatique

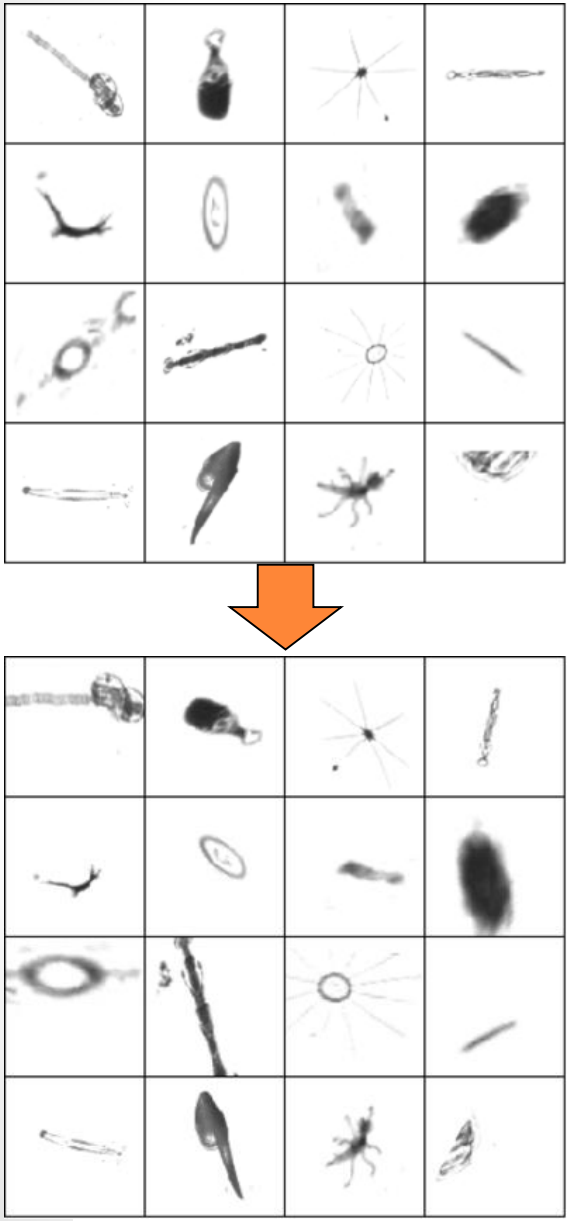
Augmenter ses données

Jean-Claude Houbart

Augmentation de données

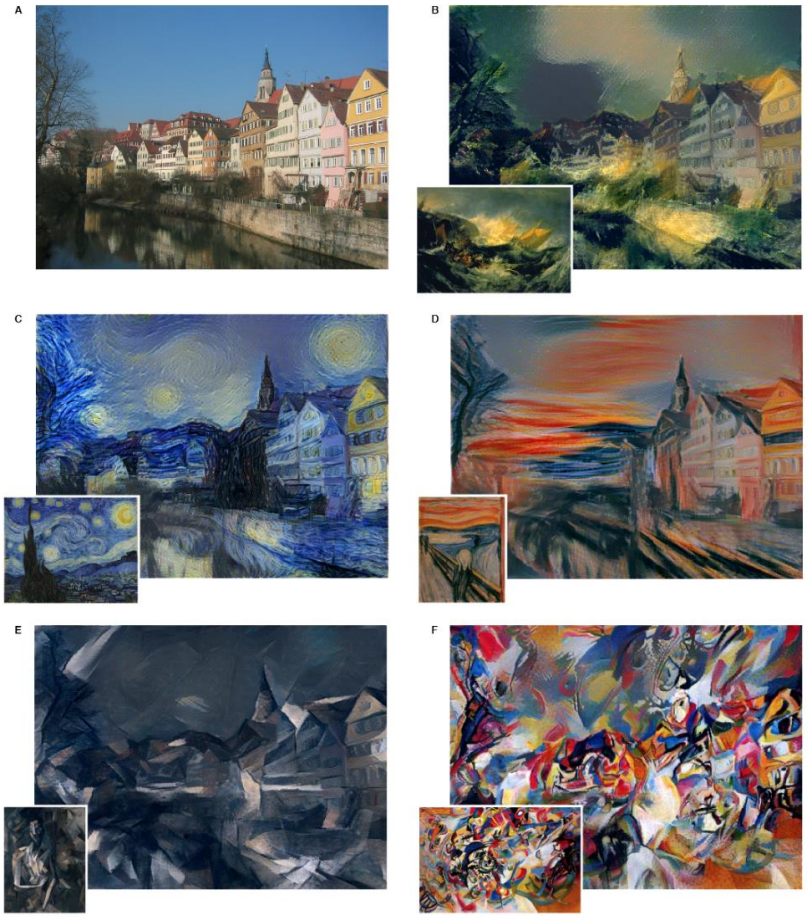
- On peut créer des données fictives à partir des données existantes.
- Cela permet d'augmenter le nombre d'exemple et de limiter ainsi le sur-apprentissage.
- Il faut faire attention à ne pas créer des biais et fausser les prédictions en inférence.

Génération d'images



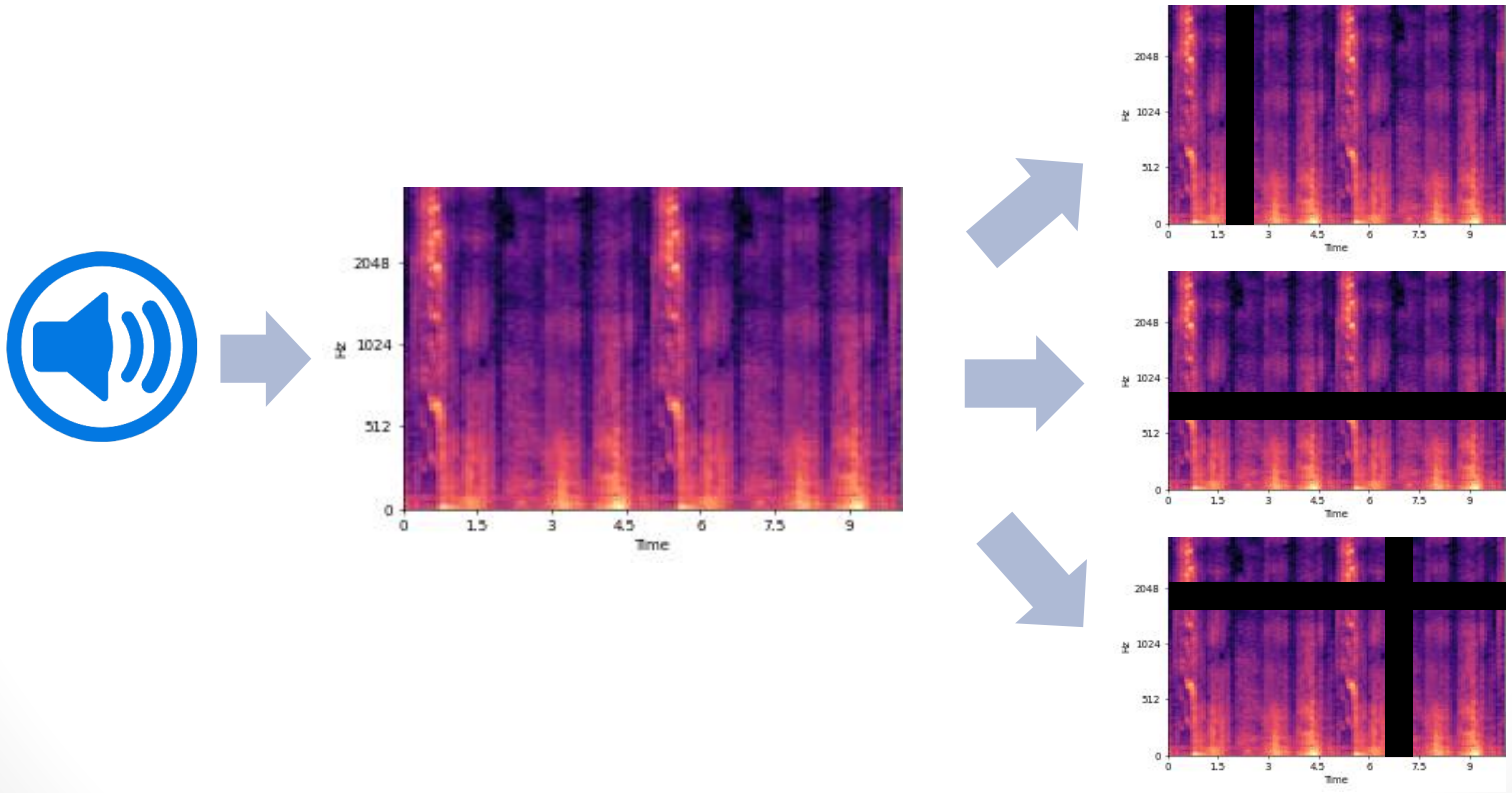
Augmented image

- On peut créer des nouvelles images en déformant les images existantes : contraste ; taille ; rotation …
- Neural Style Transfer : appliquer un style à une image



Génération de sons

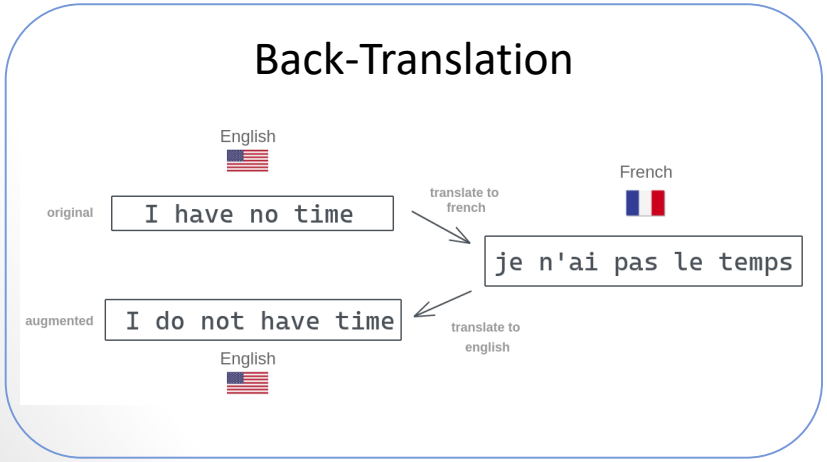
- A partir d'un son, on génère un spectrogramme : distribution des fréquences dans le temps.
- On peut alors créer de nouveaux spectrogram en masquant aléatoirement une plage de fréquence ou de temps, ou les deux.
- Les sons peuvent alors être traités comme des images.

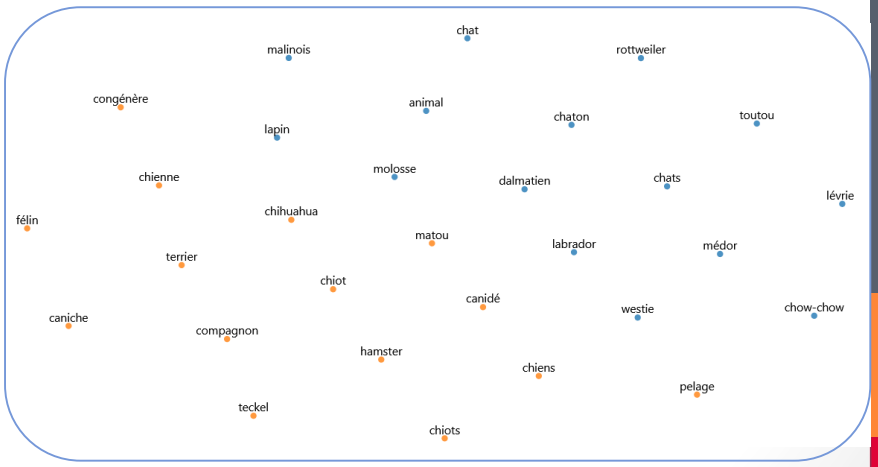


Génération de textes

- Différentes opérations permettent de générer des nouvelle phrases, même si elles n'ont pas de sens.

Opération	Exemple
Échange aléatoire	Ce chien aboie le soir => Ce chien salade le soir
Insertion aléatoire	Ce chien aboie le soir => Ce chien salade aboie le soir
Suppression aléatoire	Ce chien aboie le soir => Ce chien le soir
Synonyme aléatoire	Ce chien aboie le soir => Ce chien jappe le soir
Mot proche aléatoire	Ce chien aboie le soir => Ce labrador aboie le soir





EPSI l'École d'ingénierie informatique

Trouver des données

Jean-Claude Houbart

Collecte de données

- Il arrive souvent que les données disponibles sont en quantité insuffisante, ou que la labellisation demande trop de temps.
- Trois types d'approche sont possibles, et peuvent être combinées:
  - Trouver des datasets utiles pour notre problématique.
  - Récupérer automatiquement des données sur Internet (scrapping).
  - Faire labéliser ses données par un service spécialisé.
- Enfin, un nouveau domaine émergent est celui des données synthétiques.

Recherche de Datasets

- Il y a beaucoup de datasets disponibles sur Internet. Il faut soigneusement vérifier les droits associés.
- Les principaux frameworks de Machine Learning intègre des datasets.

  https://pytorch.org/vision/stable/datasets.html
  https://torchtext.readthedocs.io/en/latest/datasets.html
  https://scikit-learn.org/stable/datasets.html

- Le site Kaggle, déjà présenté, intègre une section:

  https://www.kaggle.com/datasets
- Google propose un moteur de recherche dédié aux datasets:

  https://datasetsearch.research.google.com/
- Et aussi des vidéos YouTube annotées :

  https://research.google.com/youtube8m/
- Le site paperswithcode recense tous les datasets utilisés et les articles scientifiques associés.

  https://paperswithcode.com/datasets/
- Enfin, le site data.world est un catalogues de dataset récentes ou plus anciennes.

  https://data.world

Scrapping

- Il existe de nombreux outils et librairies pour récupérer des données sur Internet.
- Récupération d'images. Par exemple, la librairie simple_image_download permet de charger des images selon un mot clé.
- Récupération de messages sur les réseaux sociaux, par exemple avec la librairie SNScrape.
- Récupération d'information dans des pages web avec deux étapes :
  - Récupération des pages selon des critères.
  - Analyse du contenu de la page pour extraire les informations.
  - Plusieurs solutions (généralement payantes) sont disponibles:
    - Octoparse
    - ScrapingBee
    - ScrapingBot
    - …

Labellisation

- Il est possible de faire labelliser des données pour des budgets limités.
- Amazon Mechanical Turk
- Clickworker
- CloudCrowd

- Suivant la complexité de la labellisation, le cout est de 0,1 à quelques dollars par label.