Une donnée eticté est une donné labelisé (ex, une image taggé « chat » pour une photo de chat, ou la catégorisation d’un commentaire comme comentaire « triste » ou « elogieux »)

Feature = parametre d’entrée, target = sortie

Le clustering est un regrouppement, ne pas confondre classification et clustering (classifier et regroupper)

Clustering c’est rechercher une corrélation entre des données pour les regroupers entre 2/3/5 plusieurs groupes. Clustering avec apprent non suppervisé, et classification avec apprent suppervisé.

Difference entre datascience et analyse de donnée est la presence de l’aprentissage automatique.

Outils pour la data : adoop, spark, DATAQ, power BI , ...

Différence entre analyse de donnée et extraction db, est que dans l’analyse de donnée l’on n’alterne pas la base de donné exploité. On realise des traitement pour visualiser la corélation entre des données.

Ocr :

En python il existe plusieurs librairies ppur l’exploration de données, comme pandas.

Pour la création d’interfaces d’interprétation graphiques : matplotlib, seaborn, ...

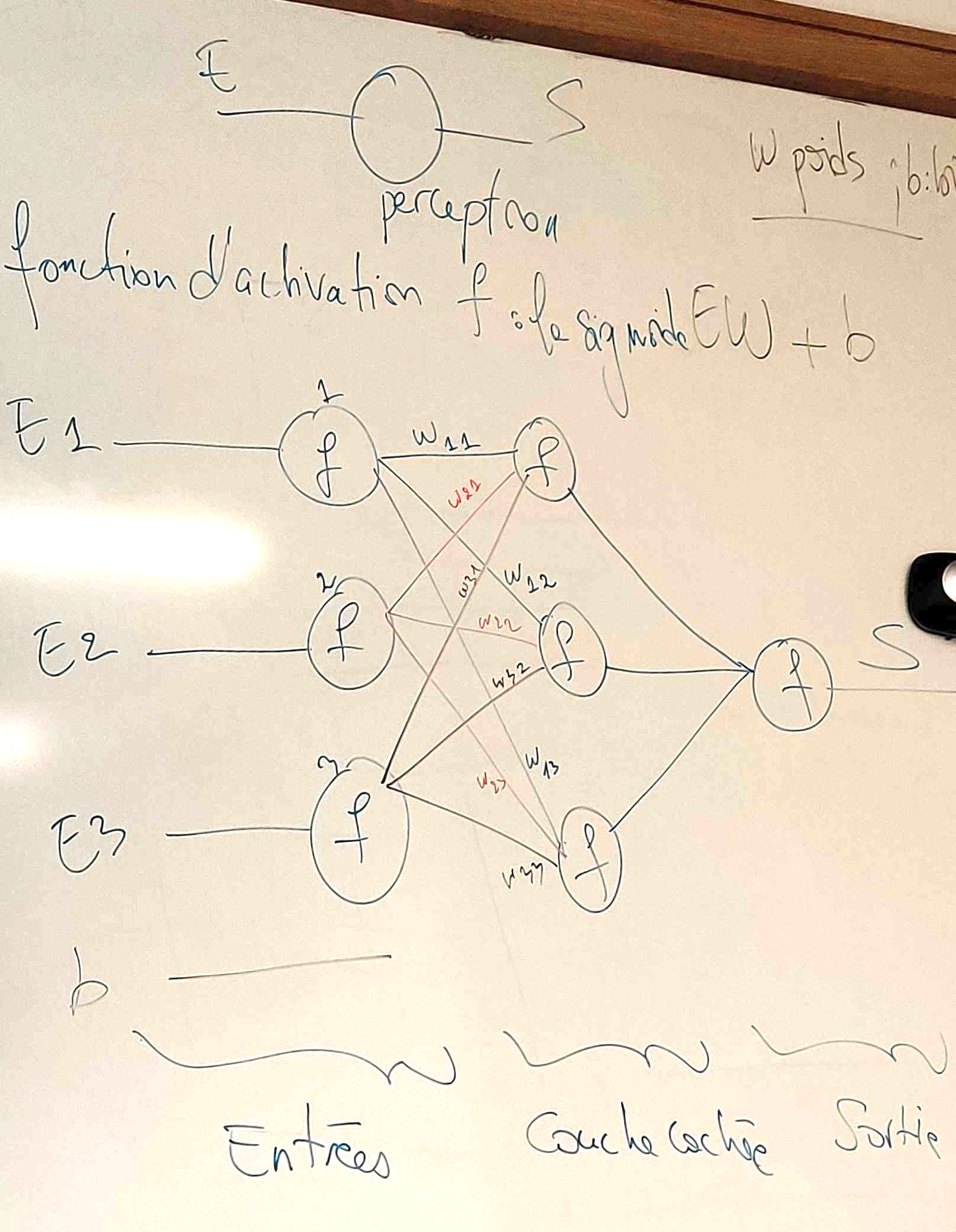
Risque d’utiliser des jeux de données non netoyer, on obtie des données erronés. Dans une demarche « green IT » utiliser des jeux de données triés et propres permet une phase d’aprentissage plus courte et don moins consomatrice en ressources.

Algo non supervisés : algorythmie génétique,

Contrairement à l’aprentissage supervié, le non supervisé ne dispose pas d’enseignant ou de contraintre. Il est libre de ses interpretations.

Matrice de confusion : on a les classes prédite et les classes reelles. C’est un outil courament utilisé pour evaluer les performances d’un modele de machine learning.

Epoque = nombre de phase d’apprentissage. Aprentissage se fait sur plusieurs iterations. Dans une itération il y a plusieurs epoques.54



Fonction sigmoide

Représentation graphique de la fonction sigmoide (par gpt) :  
