

Introduction à l'apprentissage statistique

Par Espéran Padonou

Ecole d'Eté sur
**L'INTELLIGENCE
ARTIFICIELLE**

Godomey, le 29 juillet 2024



Définition ?

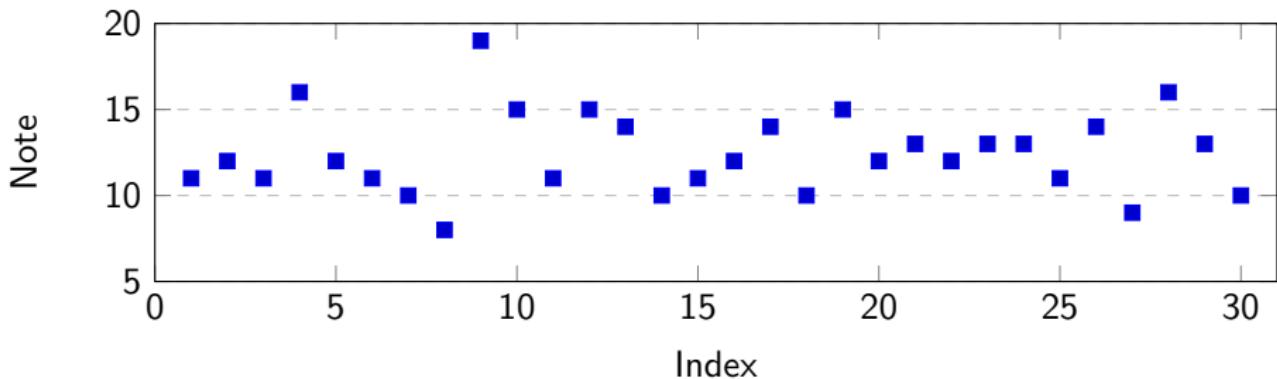
- Apprentissage automatique
- Apprentissage artificiel
- Apprentissage statistique
- Statistical learning
- Simulation de l'intelligence humaine
- Ou d'une certaine intelligence animale ?

Ingrédients: Mathématiques, statistiques, probabilités, algorithmique.

Problème 1 : détecter un événement exceptionnel

11	12	11	16	12	11	10	8	19	15	11	15	14	10	11
12	14	10	15	12	13	12	13	13	11	14	9	16	13	10

Représentation des notes des élèves



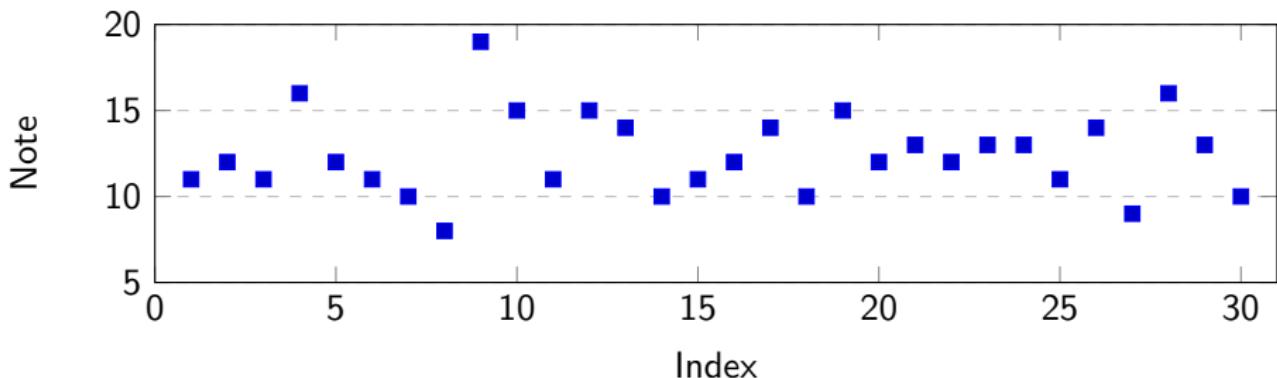
Y a-t-il une note exceptionnelle ?

- Comment la définissez-vous ?

Problème 1 : détecter un événement exceptionnel

11	12	11	16	12	11	10	8	19	15	11	15	14	10	11
12	14	10	15	12	13	12	13	13	11	14	9	16	13	10

Représentation des notes des élèves



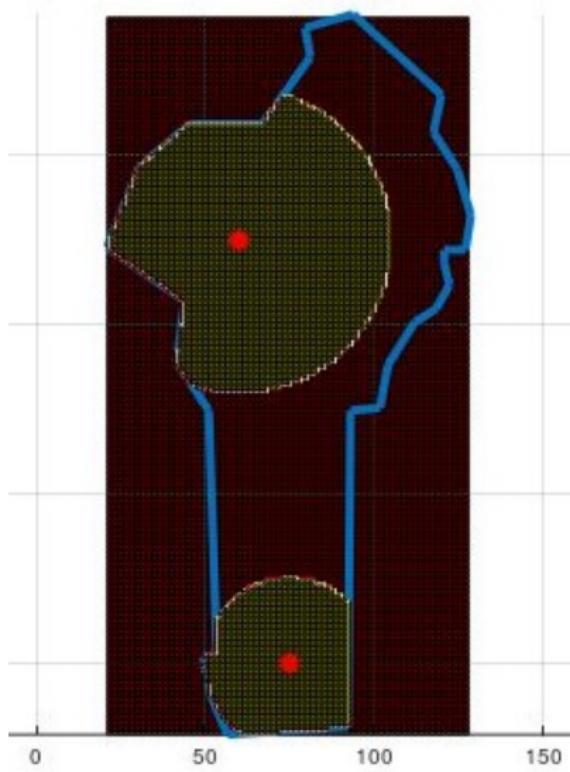
Y a-t-il une note exceptionnelle ?

- Comment la définissez-vous ?
- Comment un ordinateur peut-il la retrouver ? Règle, formule ?

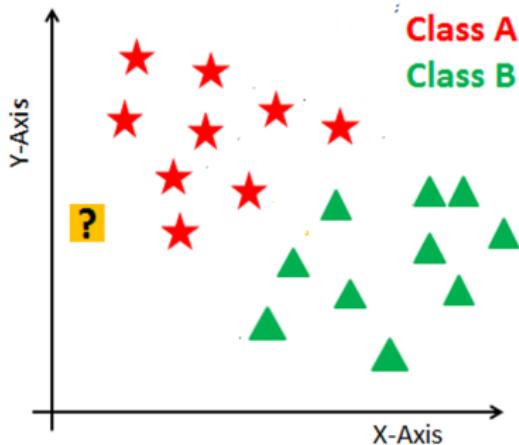
Problème 2 : optimisation

Où positionner une antenne

- Avec quel objectif ?
- Quelles variables ?
- Quelle stratégie ?
- Quelles contraintes ?



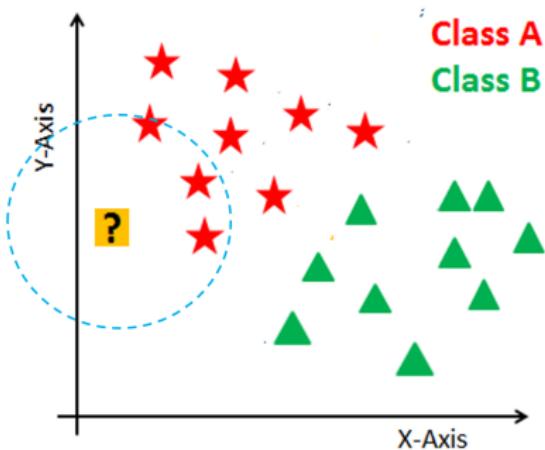
Problème 3 : classification



Trouver la couleur du nouveau point

- En sachant sa position
- En sachant les couleurs des anciens points

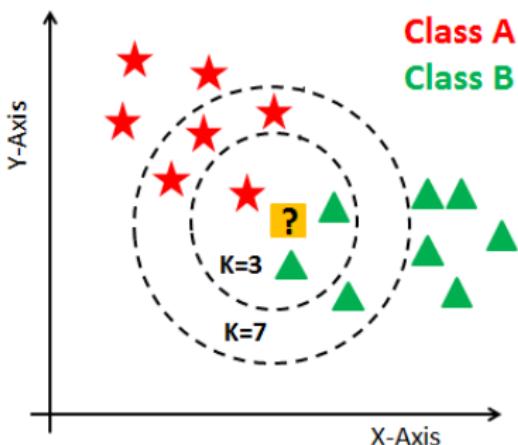
Problème 3 : classification



Algorithme des k plus proches voisins

- Calculer des distances
- Les classer par ordre croissant
- Retenir les k plus petites valeurs et compter

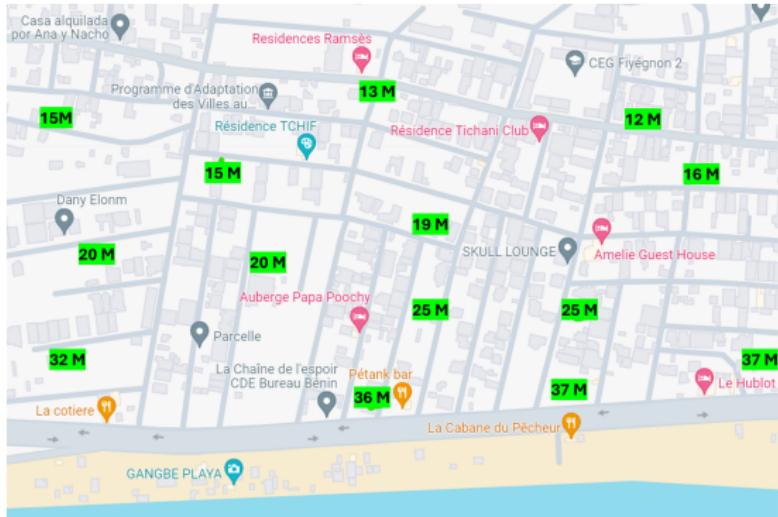
Problème 3 : choisir k



Vers la sélection de modèles

- Comment évaluer le cas $k = 3$?
- Des idées...

Problème 3 : exercice



Prix de vente de terrains dans une zone convoitée de Cotonou

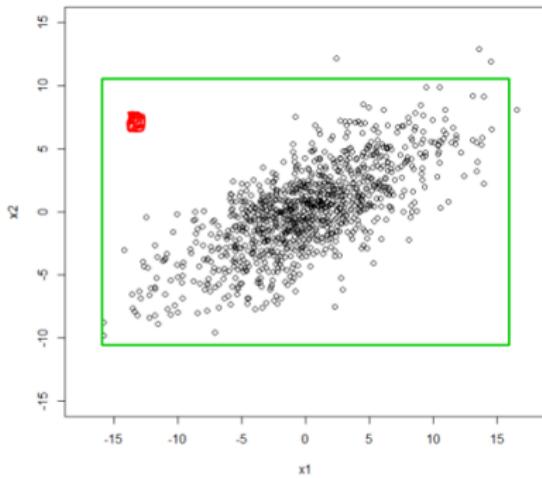
- Modifiez l'algorithme (sortie quantitative et même aire).
- Variations horizontales VS variations verticales : commentez. Idées ?
- Idées si les surfaces sont différentes ? Pour différentes années de vente.

Contrôle de gestion

- Quantité de canne à sucre achetée (tonnes) : $MP \in [0, 30]$
- Quantité de sucre fabriqué (caisses) : $PF \in [0, 30]$
- Redressement fiscal bien que MP et PF soient dans leur intervalle historique

Contrôle de gestion

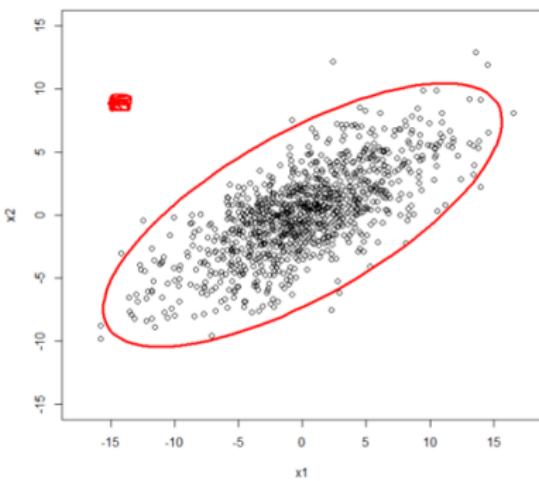
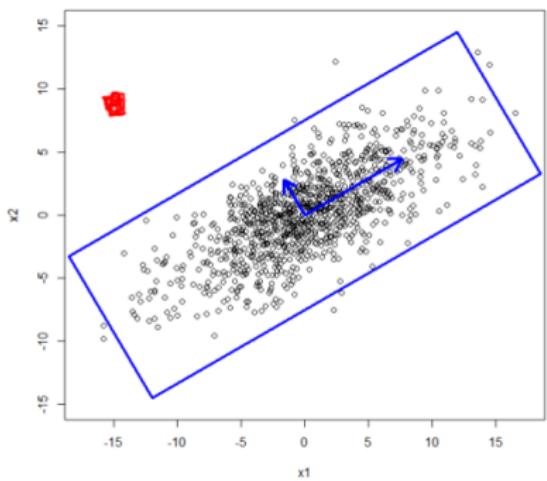
- Quantité de canne à sucre achetée (tonnes) : $MP \in [0, 30]$
- Quantité de sucre fabriqué (caisses) : $PF \in [0, 30]$
- Redressement fiscal bien que MP et PF soient dans leur intervalle historique



Problème 4 : l'analyse multivariée

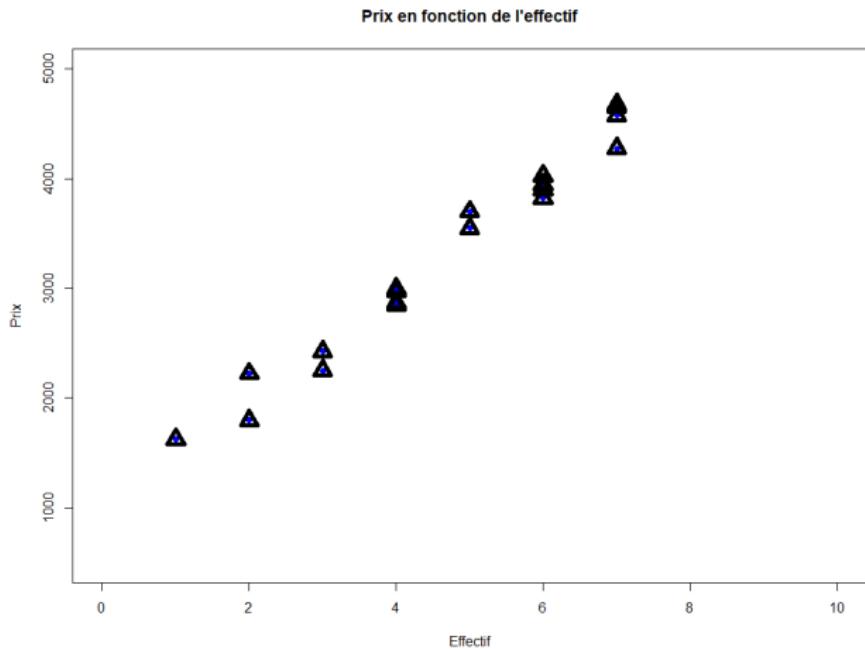
Contrôle de gestion

- Grand axe : ordre de grandeur de MP et PF
- Petit axe : lien entre MP et PF
- La relation entre les variables devient un indicateur à surveiller
- Vers l'Analyse en Composantes Principales



Problème 5 : Régression

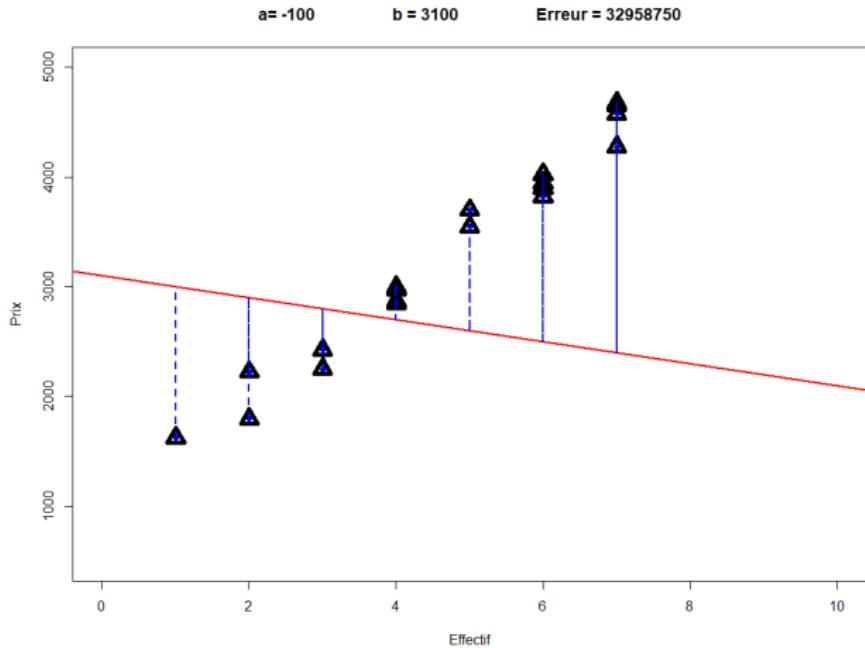
Effectif	Prix
4	3000
3	2250
3	2425
1	1625
2	2225
7	4275
6	4025
6	3950
7	4575
4	2875
4	2975
6	3825
5	3700
7	4675
6	3900
2	1800
5	3550
7	4650
4	2850



- Que dire des prix pratiqués par ce coiffeur ?
- Quel est le prix probable pour 10 personnes ?

Problème 5 : Régression

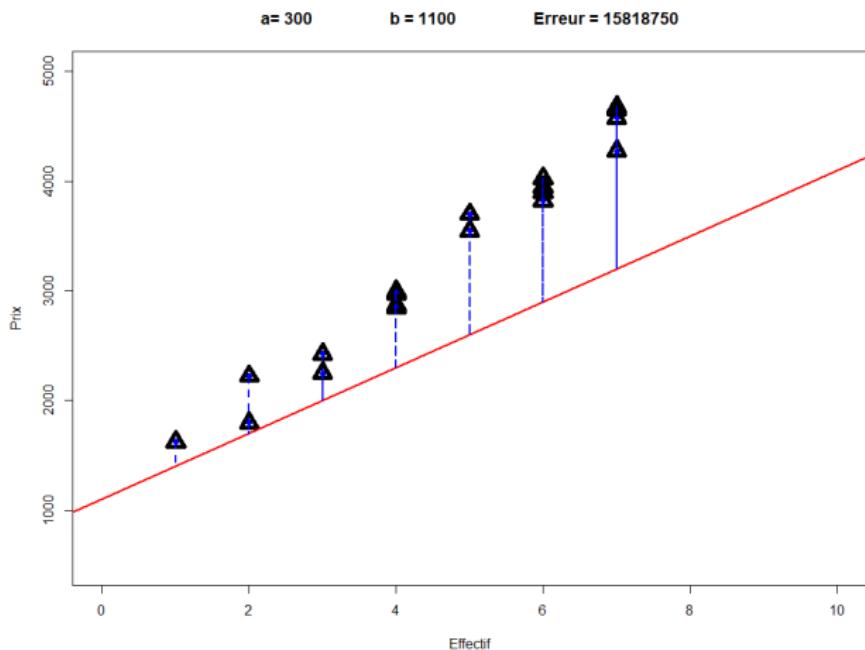
Effectif	Prix
4	3000
3	2250
3	2425
1	1625
2	2225
7	4275
6	4025
6	3950
7	4575
4	2875
4	2975
6	3825
5	3700
7	4675
6	3900
2	1800
5	3550
7	4650
4	2850



- Définir les moindres carrés
- Commentez le choix $a = -100$

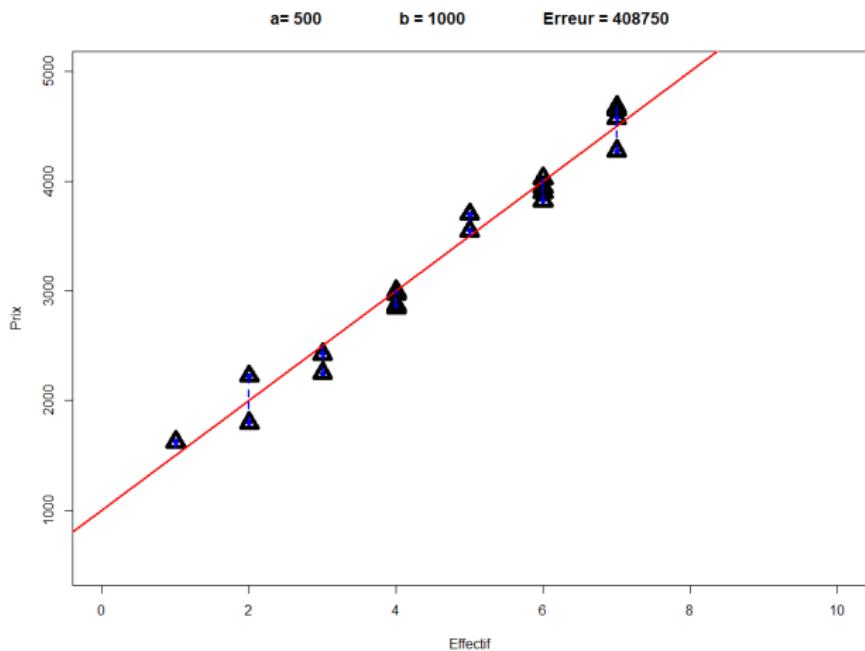
Problème 5 : Régression

Effectif	Prix
4	3000
3	2250
3	2425
1	1625
2	2225
7	4275
6	4025
6	3950
7	4575
4	2875
4	2975
6	3825
5	3700
7	4675
6	3900
2	1800
5	3550
7	4650
4	2850



Problème 5 : Régression

Effectif	Prix
4	3000
3	2250
3	2425
1	1625
2	2225
7	4275
6	4025
6	3950
7	4575
4	2875
4	2975
6	3825
5	3700
7	4675
6	3900
2	1800
5	3550
7	4650
4	2850



Problème 6 : Paradoxe de Simpson

	année 2021		année 2022	
	inscrits	reçus	inscrits	reçus
non redoublants	22	12	15	8
redoublants	3	3	10	9

Deux commentaires qui s'opposent

- Le directeur : Notre taux de réussite a augmenté de plus de 13% !
- Toto : Les redoublants ont moins réussi cette année.
Les non-redoublants également ont moins réussi qu'en 2021!
- Vers les modèles par arbre : un modèle pour les redoublants et un modèle pour les autres.

Problème 6 : Paradoxe de Simpson

	année 2021		année 2022	
	inscrits	reçus	inscrits	reçus
non redoublants	22	12	15	8
redoublants	3	3	10	9

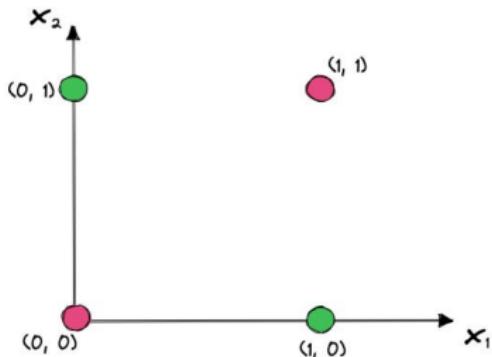
Deux commentaires qui s'opposent

- Le directeur : Notre taux de réussite a augmenté de plus de 13% !
- Toto : Les redoublants ont moins réussi cette année.
Les non-redoublants également ont moins réussi qu'en 2021!
- Vers les modèles par arbre : un modèle pour les redoublants et un modèle pour les autres.

Problème 7 : un exemple classique

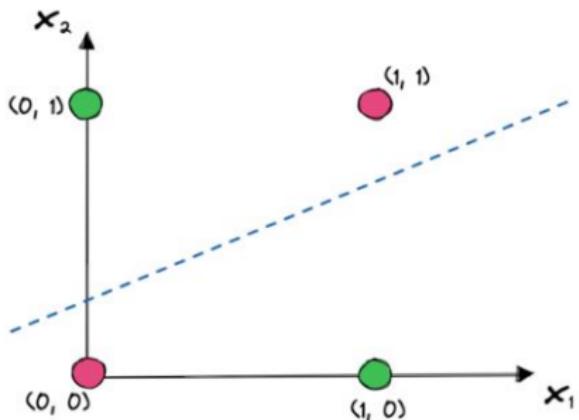
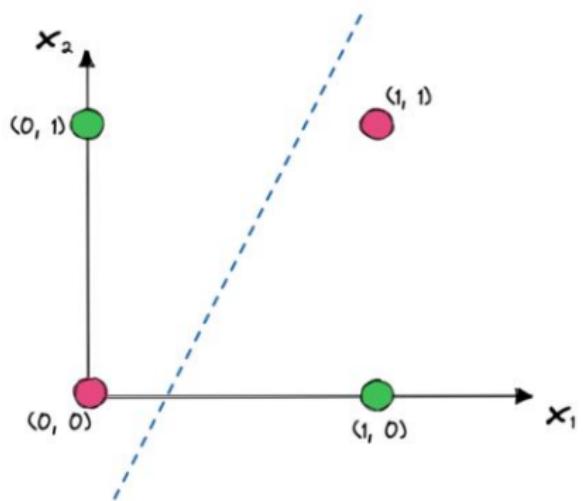
- x_1 et x_2 sont binaires et $y = x_1 \text{XOR} x_2$
- Représentation de y en code couleur dans le plan (x_1, x_2)

x_1	x_2	$x_1 \text{ XOR } x_2$
0	0	0
0	1	1
1	0	1
1	1	0

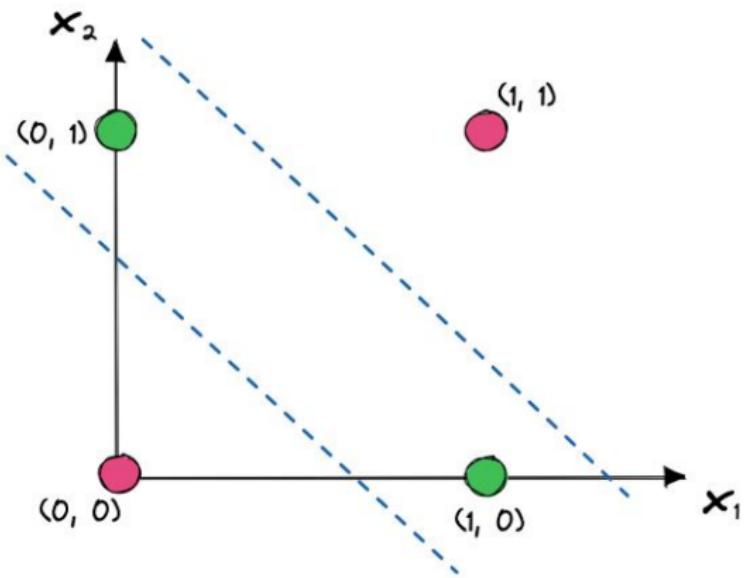


- Comment prédire la couleur de l'un des points s'il était inconnu ?
 - Par les k plus proches voisins ?
 - Par régression linéaire ?

Problème 7 : un exemple classique

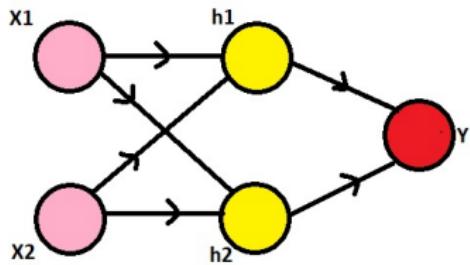
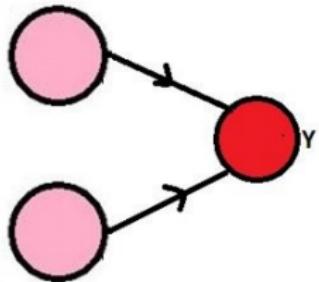


Problème 7 : un exemple classique



- Première ligne pour vérifier qu'au moins une valeur vaut 1
- Seconde ligne pour vérifier que x_1 et x_2 ne valent pas 1 à la fois

Problème 7 : Vers les réseaux de neurones



- Introduire dans le modèle de régression des variables intermédiaires, combinant les entrées et influençant la sortie, invisibles à l'entrée et à la sortie du modèle
- Un pas vers l'apprentissage profond et les réseaux de neurones

Problème 8 : Planification d'expériences



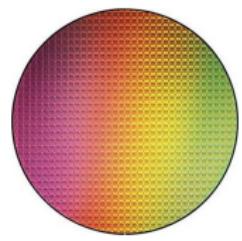
Connected objects



Integrated circuits

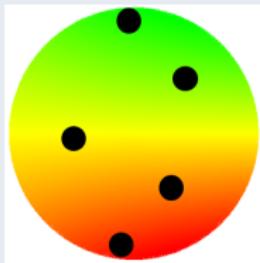
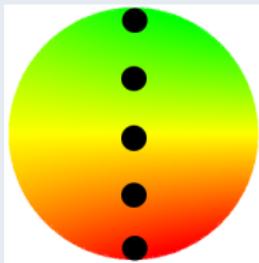
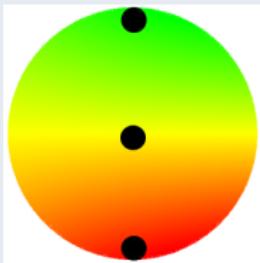
Contexte industriel

- Travail en partenariat avec STMicroelectronics.
 - ⇒ Production de circuits intégrés.
 - ⇒ Supports d'usinage en forme de disques.
- Haute qualité et contrôles coûteux.
 - ⇒ IA pour le Contrôle de Avancé de Procédés

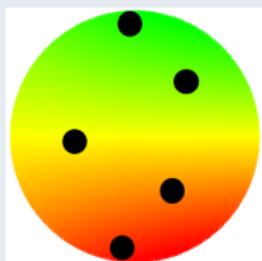
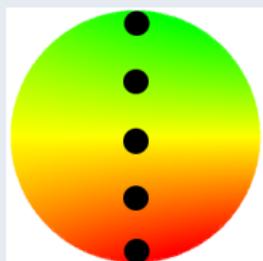
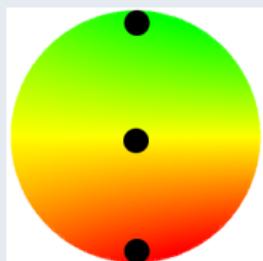
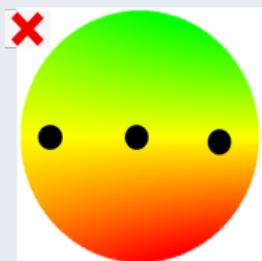


Un wafer

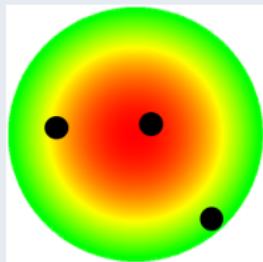
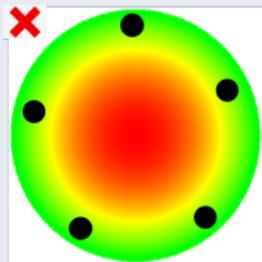
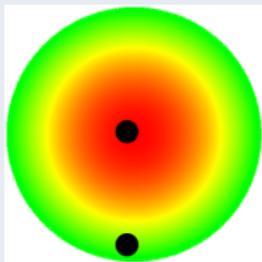
Plans d'expériences adaptés au procédé laser



Plans d'expériences adaptés au procédé laser

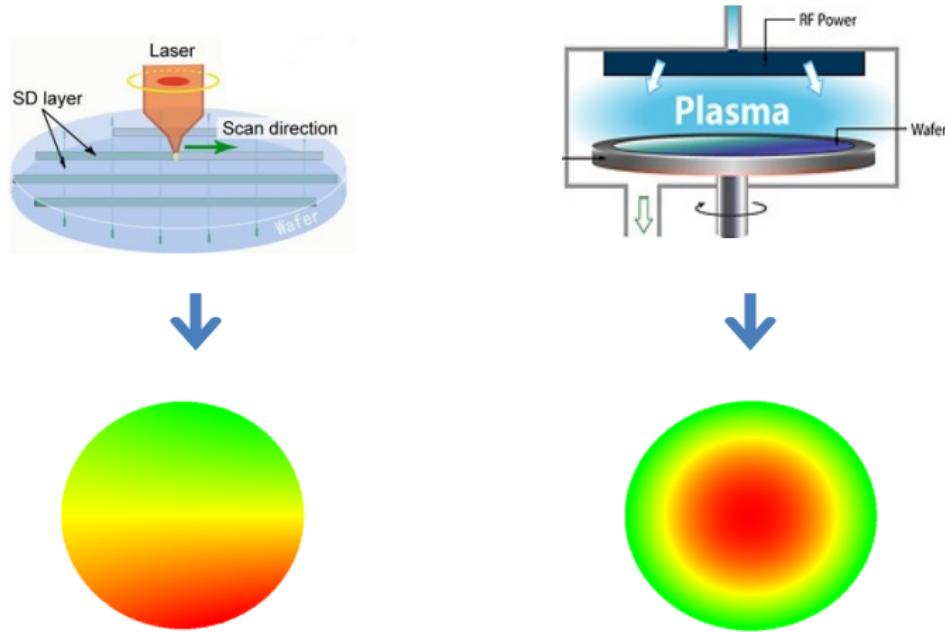


Plans d'expériences adaptés au dépôt en phase plasma



La qualité des données plutôt que la quantité.

Problème 9 : interprétation et connaissance métier



Grande variété de formes, due aux procédés d'usinage

- ▶ Les données, mais aussi les processus sous-jacents.

Problème 10 : les outils graphiques

Il ne s'agit pas d'un problème, mais d'une clé!