
IA et Deep Learning

Introduction

Informations

- **Organisation**
 - *2 heures de cours*
 - *23 heures de TD sur NoteBook + séances pers*
 - **Evaluation**
 - *1 QCM (vendredi)*
 - *1 Projet (à rendre à une date à déterminer)*
 - **Prérequis**
 - *Algèbre linéaire*
 - *Python (apprentissage « on the fly »)*
-

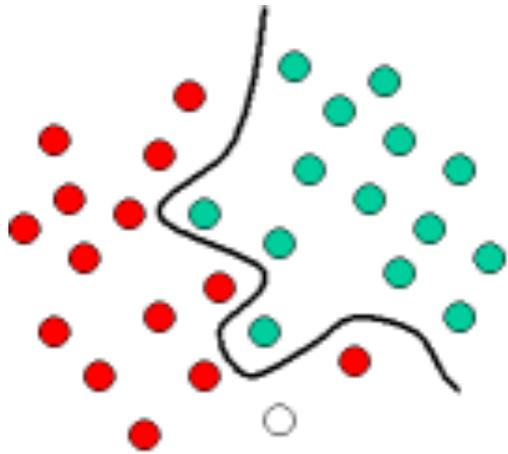
Machine Learning

Apprentissage automatique

- **Objectif** → Apprendre de la connaissance à partir de **données**

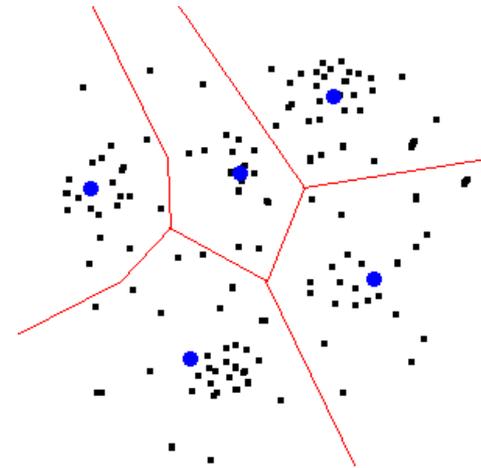
Apprentissage supervisé

Données = observations + étiquettes
Connaissance → relation entrée-sortie



Apprentissage non-supervisé

Données = observations
Connaissance → structures latentes



Apprentissage supervisé (1/2)

- A partir d'un certain nombre d'exemples...

input (data)	4	13	6	11	8
output (class label)	0	1	0	1	0

- ... on souhaite trouver une hypothèse qui explique les exemples.

question	7
prediction	?

Hypothèses possibles

- *Nombres paires*
- *Nombres plus grands de 10*

Apprentissage supervisé (2/2)

- En ajoutant des nouveaux exemples...

input (data)	4	13	6	11	8	2
output (class label)	0	1	0	1	0	1

- ... l'espace des hypothèses se réduit !

question	7
prediction	?

Hypothèses possibles

- ~~Nombres paires~~
- ~~Nombres plus grands de 10~~
- *Nombres premiers*

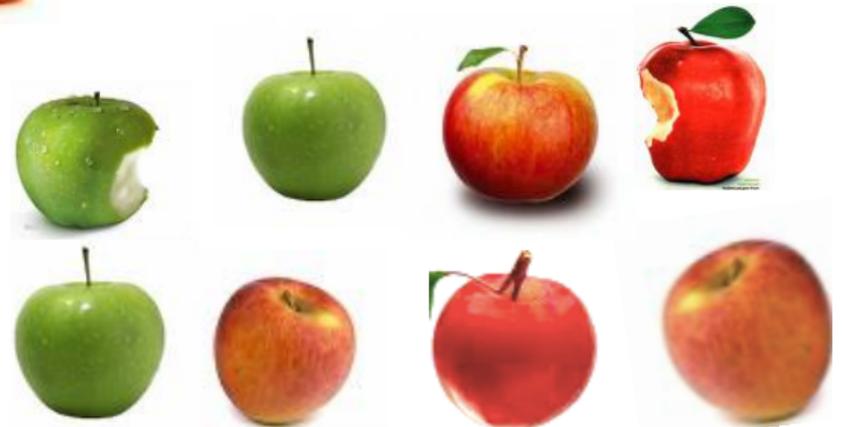
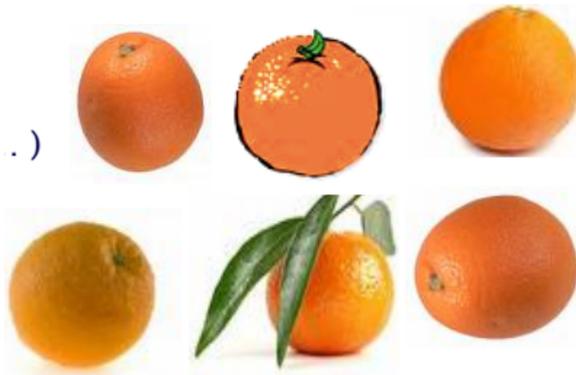
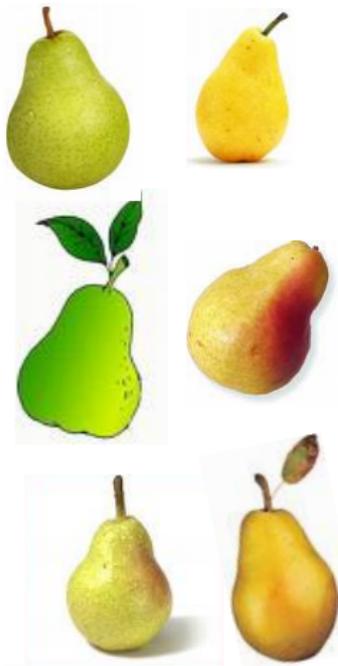
Apprentissage non-supervisé (1/3)

- A partir d'un ensemble de données (sans les réponses correctes)...



Apprentissage non-supervisé (2/3)

- ... on souhaite former des groupes respectant un critère de similarité



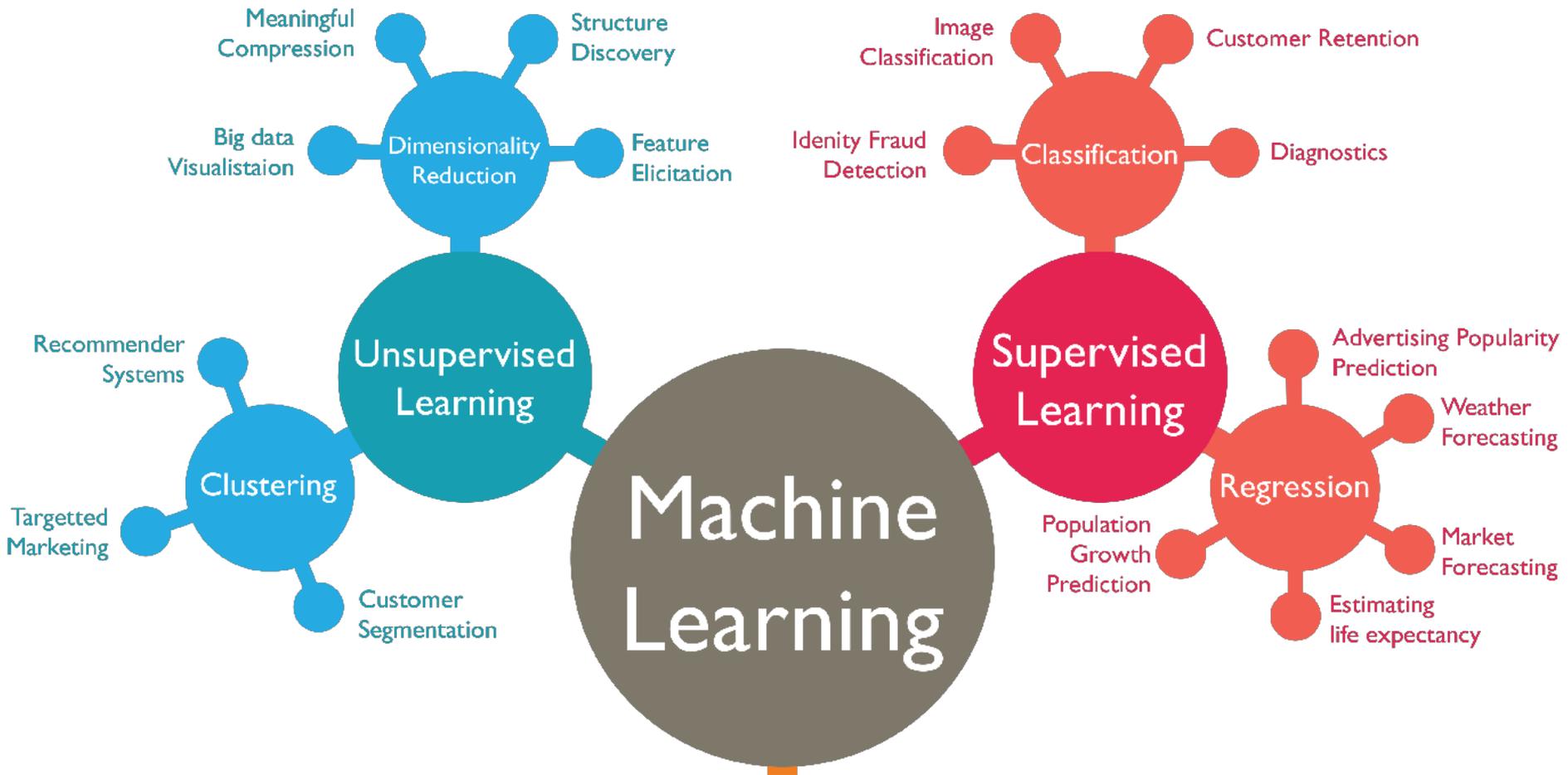
Apprentissage non-supervisé (3/3)

■ Défis

- ❑ *Quel critère de similarité ?*
- ❑ *Combien de groupes ?*



Types d'apprentissage



Un exemple d'apprentissage

Tout d'abord, les données

■ Données

- *Caractéristiques des appartements de **New York** et **San Francisco***

Altitude (m)	Année	Chambres	m2	Prix total	Prix / m2	Ville
10	2005	2	50	500'000	10'000	NY
125	1998	1	45	600'000	13'000	SF
...

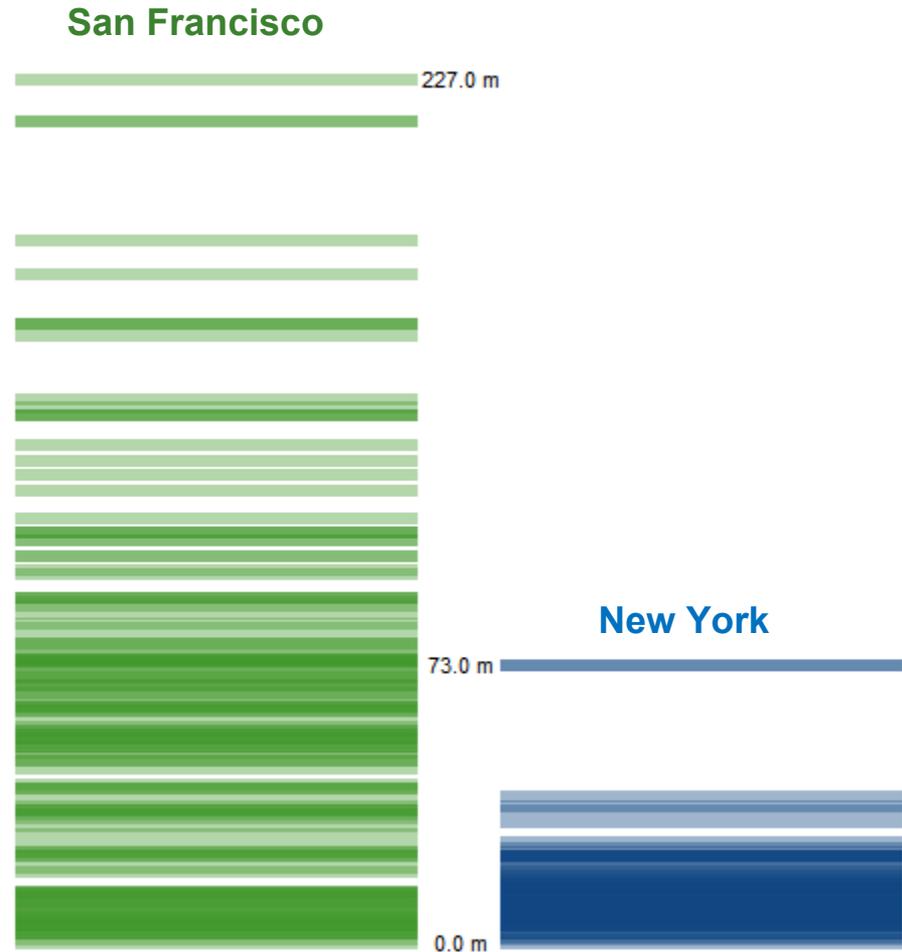
■ Problème

- *A partir des caractéristiques d'un appartement...*
- *... prédire s'il trouve à **New York** ou **San Francisco***

Un peu d'intuition

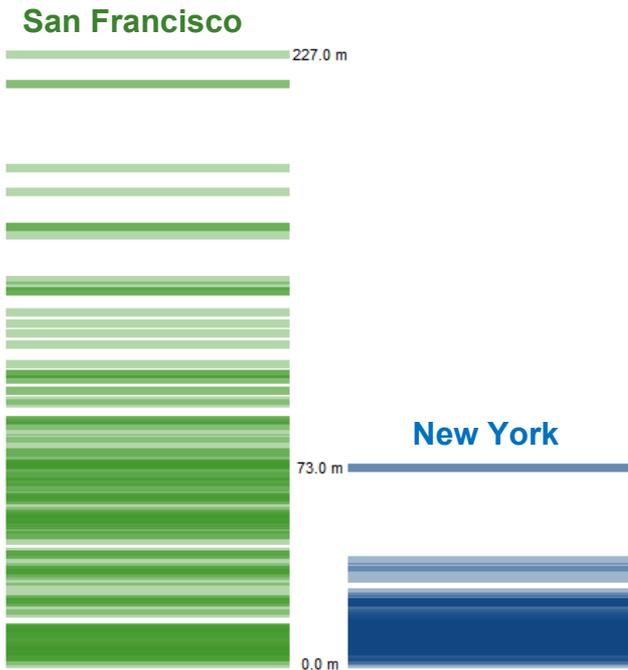
■ Première hypothèse

- *Puisque San Francisco est plutôt vallonnée, l'altitude d'un appartement semble être un bon moyen de distinguer entre les deux villes.*
- *A partir des données, on pourrait décider qu'un appartement au dessus de 73m devrait être **classée** comme située à San Francisco.*



Les erreurs de prédiction

■ Evaluation de l'hypothèse



Altitude > 73 m

(exactitude: 63%)

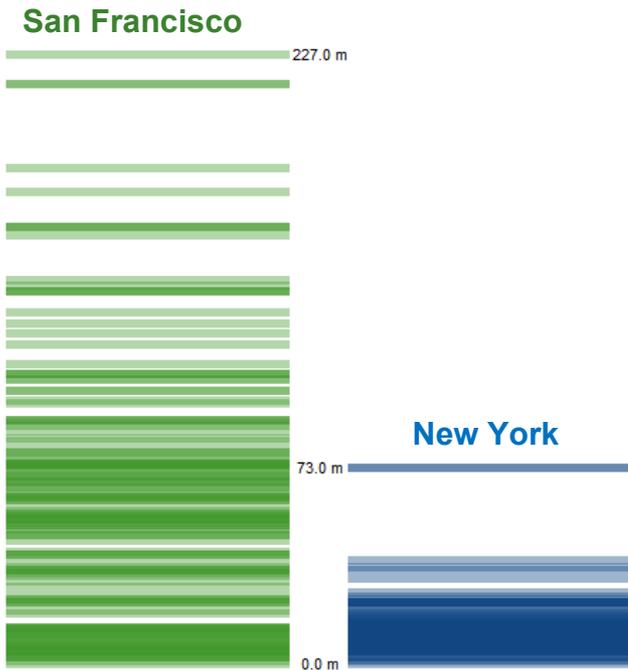
SF

NY



Le meilleur point de séparation

■ Correction de l'hypothèse



Altitude > 28 m

(exactitude: 82%)

SF

NY

(faux)

6

NY

(vrai)

99

SF

(vrai)

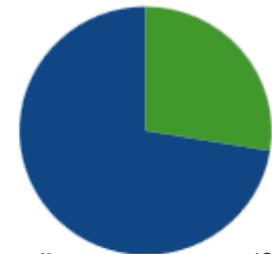
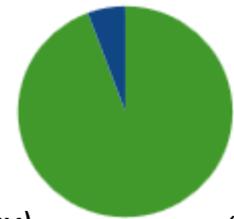
105

NY

(faux)

40

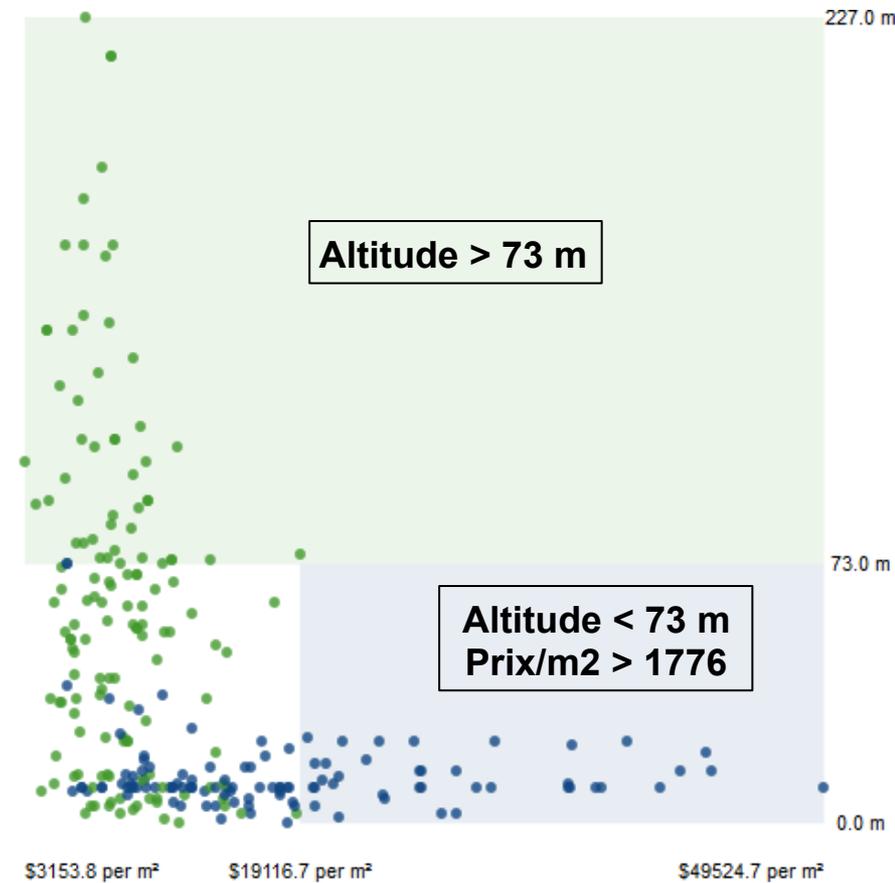
SF



Une nouvelle perspective

■ Ajout d'autres dimensions

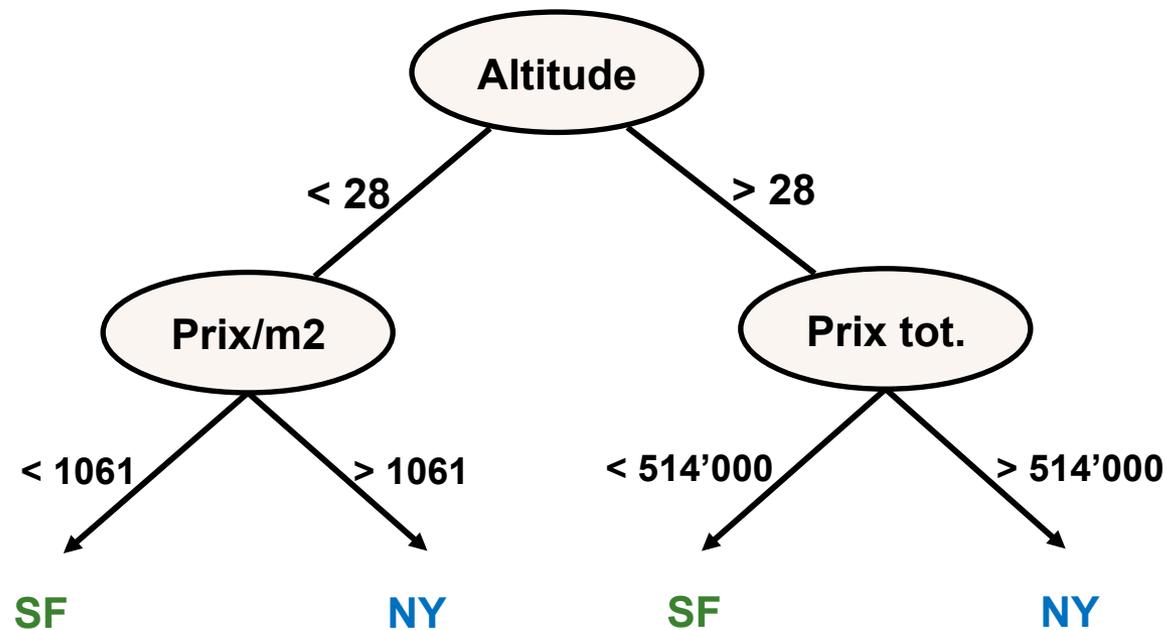
- *Ajouter d'autres informations que l'altitude permet de nuancer.*
- *Par exemple, les appartements new-yorkais peuvent atteindre des prix au mètre carré très élevés.*
- *Ainsi, en visualisant l'altitude et le prix au mètre carré avec un **nuage de points**, on peut mieux différencier parmi les maisons situées à faible altitude.*



Obtenir une meilleure prédiction

- **Nouvelle hypothèse**

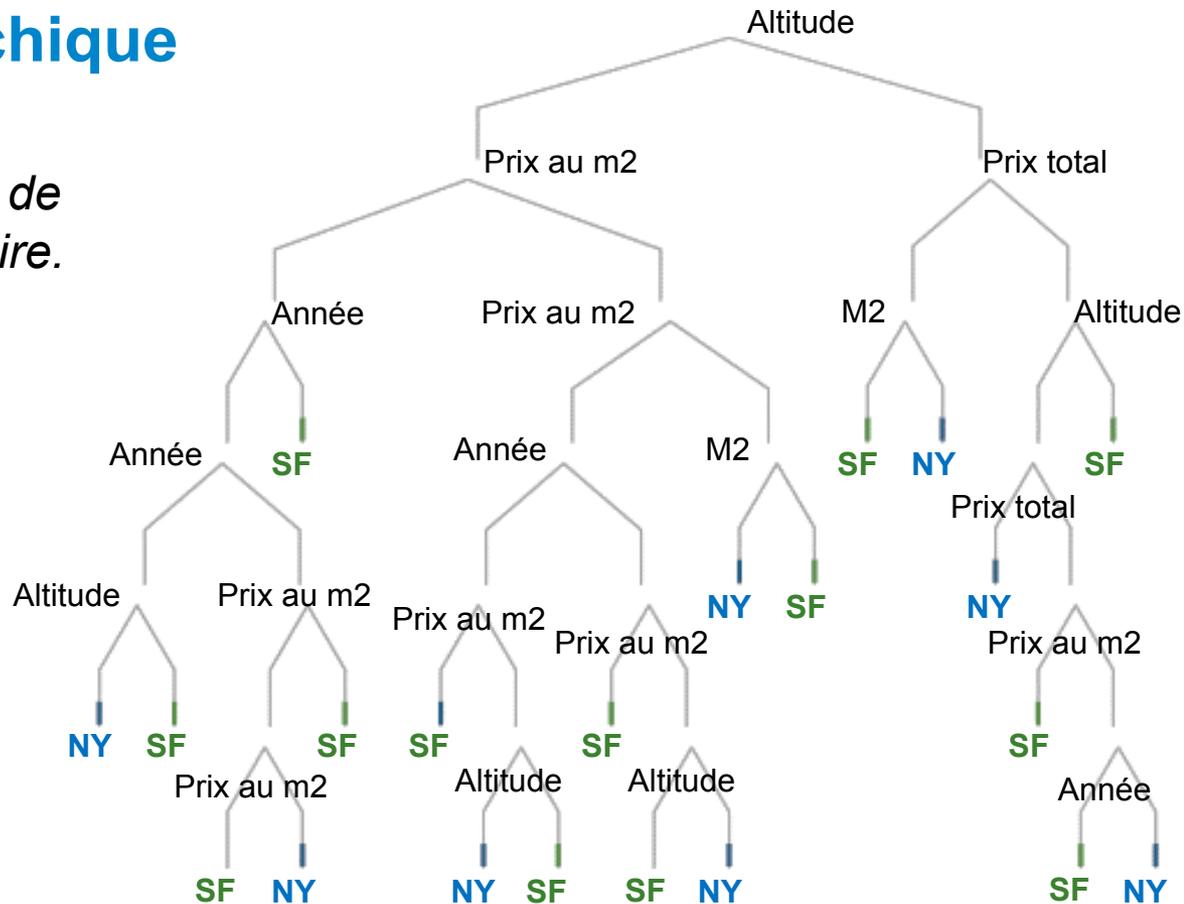
- *Des branches supplémentaires permettent d'améliorer la prédiction*



Faire pousser un arbre de décision

■ Hypothèse hiérarchique

- *On peut ajouter autant de branches que nécessaire.*
- *Les feuilles de l'arbre correspondent à une prédiction (NY ou SF)*



Données d'apprentissage

■ Prédiction

- *L'arbre de décision permet de déterminer où un appartement se trouve en parcourant les branches correspondant aux données.*
- *Ces données sont dites **d'apprentissage** car elles ont été utilisées pour apprendre le modèle. Cela est fait automatiquement par un algorithme.*
- *L'arbre ainsi obtenu permet de prédire parfaitement les données d'apprentissage, car on l'a construit sur ces données.*

précision sur les données d'apprentissage

111/111 100% 139/139



Confronter la réalité

■ Evaluation

- *Bien évidemment, ce qui importe c'est la précision de l'arbre à effectuer des prédictions pour des données jamais vues.*
- *Pour évaluer les performances de notre modèle de prédiction sur de nouvelles données, nous avons besoin de l'appliquer à des données jamais rencontrées. Ces données sont dites de **test**.*
- *La qualité d'un modèle de prédiction se mesure sur les données de test et non pas sur les données d'apprentissage.*

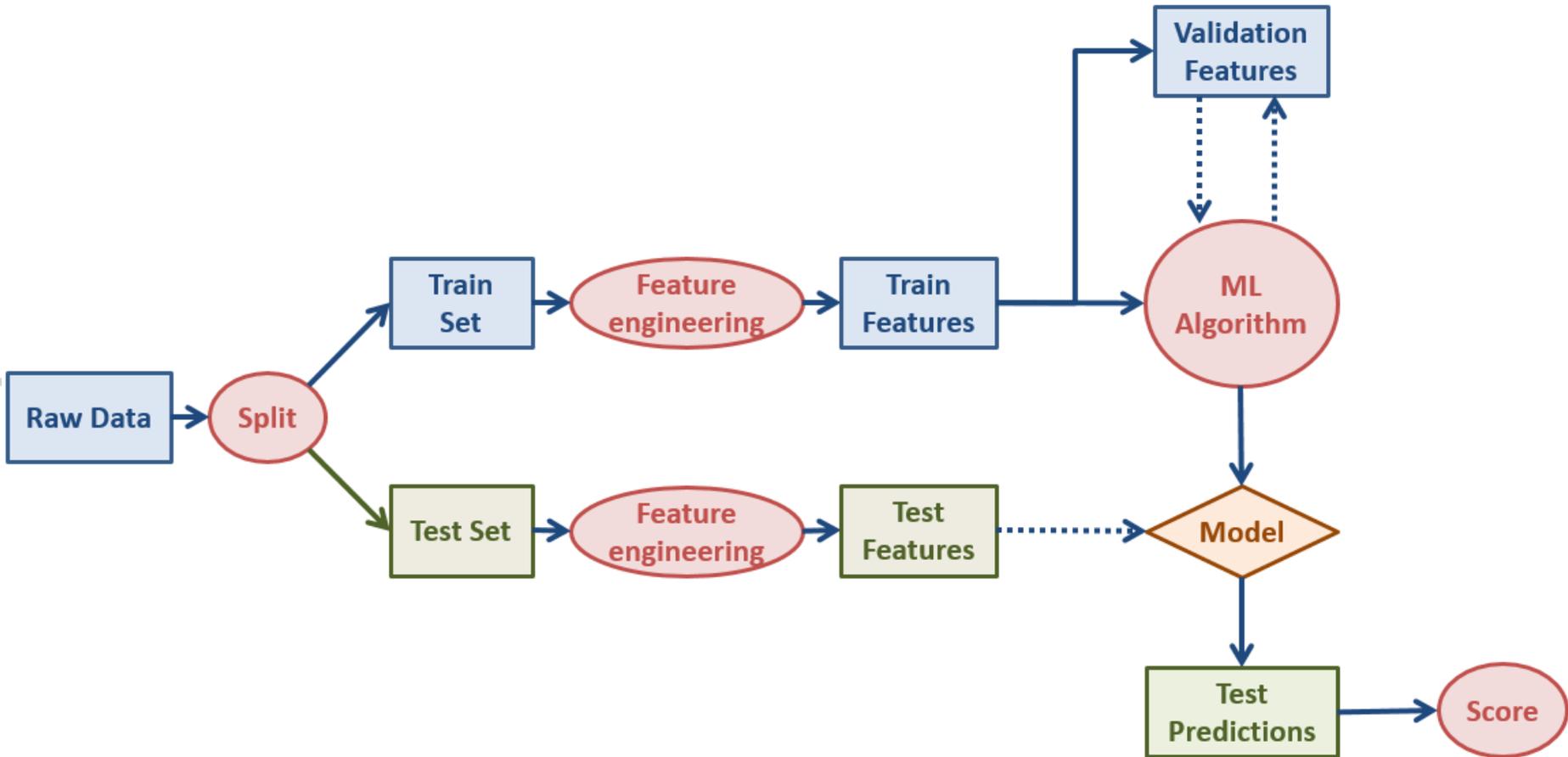


précision sur les données de test

100/112 89.7% 117/130



Schéma général d'apprentissage



Apprentissage supervisé

Données d'apprentissage

■ Contexte du probleme

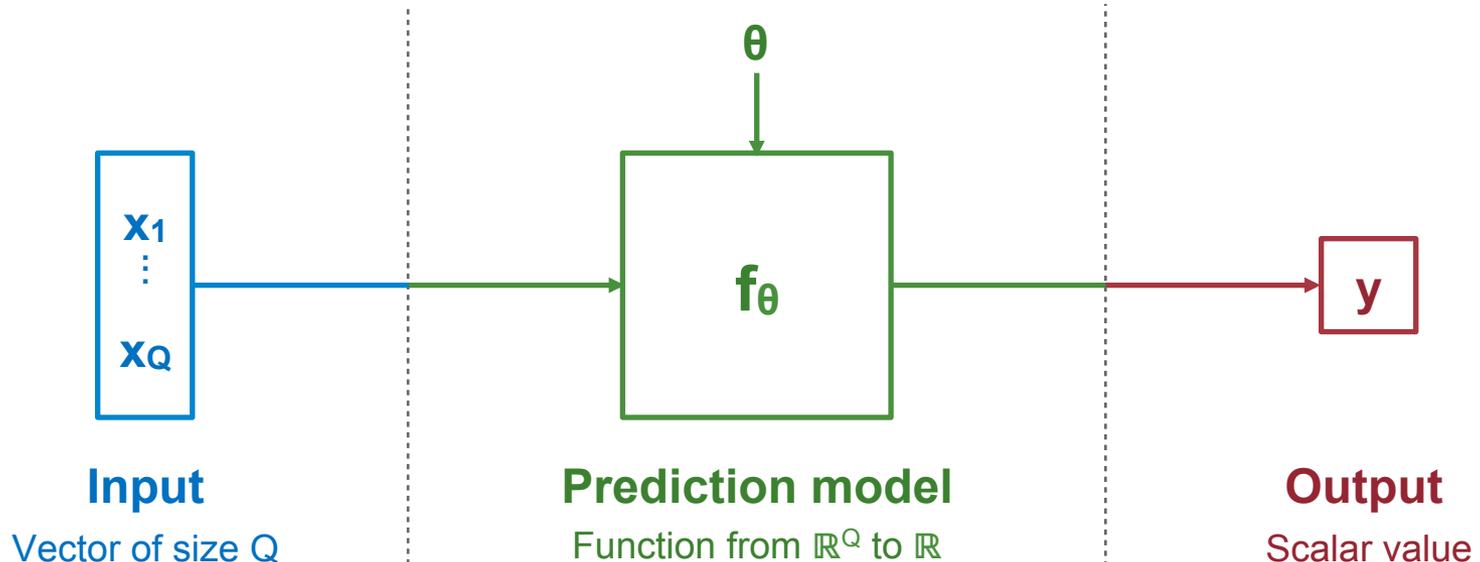
- *On veut prédire une réponse à partir de certaines observations*
- *On dispose d'exemples du type "observations - réponse correcte"*
- *On sait qu'il existe une relation entre les observations et la réponse*

	<i>Input feature 1</i>	<i>Input feature 2</i>	<i>Input feature 3</i>	<i>Input feature 4</i>	<i>Output</i>
	Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)
$(x^{(1)}, y^{(1)}) = \text{example 1}$	$x_1^{(1)} = 2104$	$x_2^{(1)} = 5$	$x_3^{(1)} = 1$	$x_4^{(1)} = 45$	$y^{(1)} = 460$
$(x^{(2)}, y^{(2)}) = \text{example 2}$	$x_1^{(2)} = 1416$	$x_2^{(2)} = 3$	$x_3^{(2)} = 2$	$x_4^{(2)} = 40$	$y^{(2)} = 232$
$(x^{(3)}, y^{(3)}) = \text{example 3}$	$x_1^{(3)} = 1534$	$x_2^{(3)} = 3$	$x_3^{(3)} = 2$	$x_4^{(3)} = 30$	$y^{(3)} = 315$
$(x^{(4)}, y^{(4)}) = \text{example 4}$	$x_1^{(4)} = 852$	$x_2^{(4)} = 2$	$x_3^{(4)} = 1$	$x_4^{(4)} = 36$	$y^{(4)} = 178$

Modèle de prediction

■ Généralisation par biais inductif

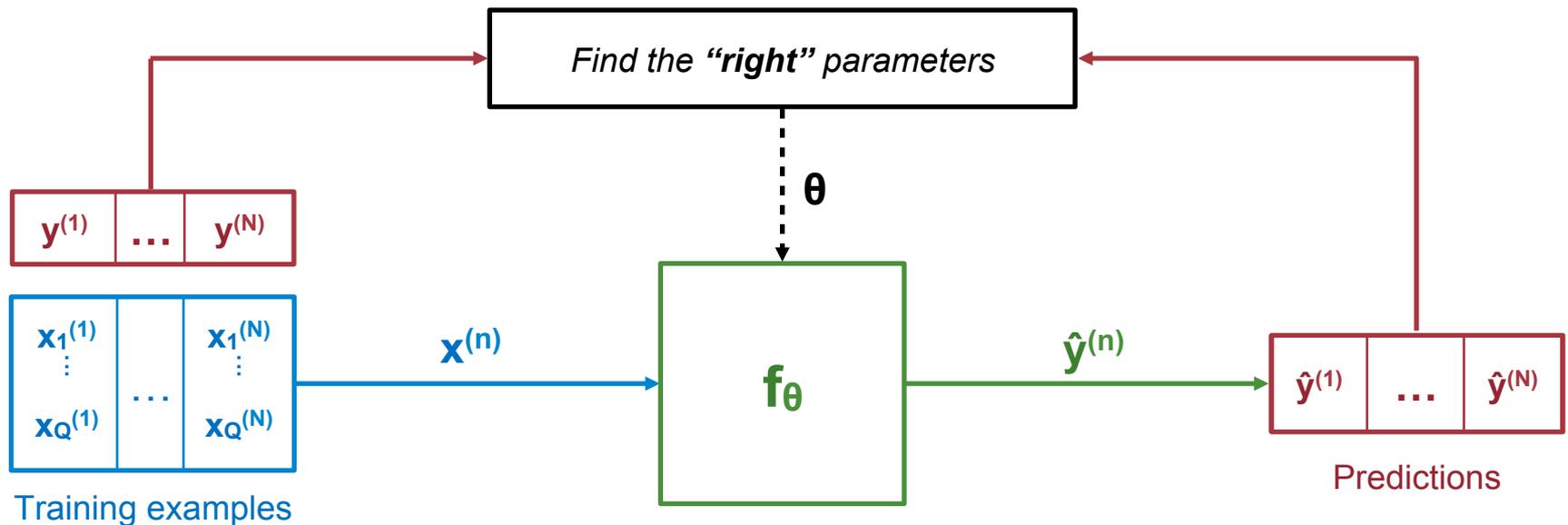
- *On est intéressé à faire des prédictions pour des nouvelles données*
- *Pour ce faire, on choisit d'abord un modèle de prediction*
- *Ce modèle est paramétré par un vecteur θ (qui est inconnu initialement)*



Méthode d'apprentissage

■ Apprentissage

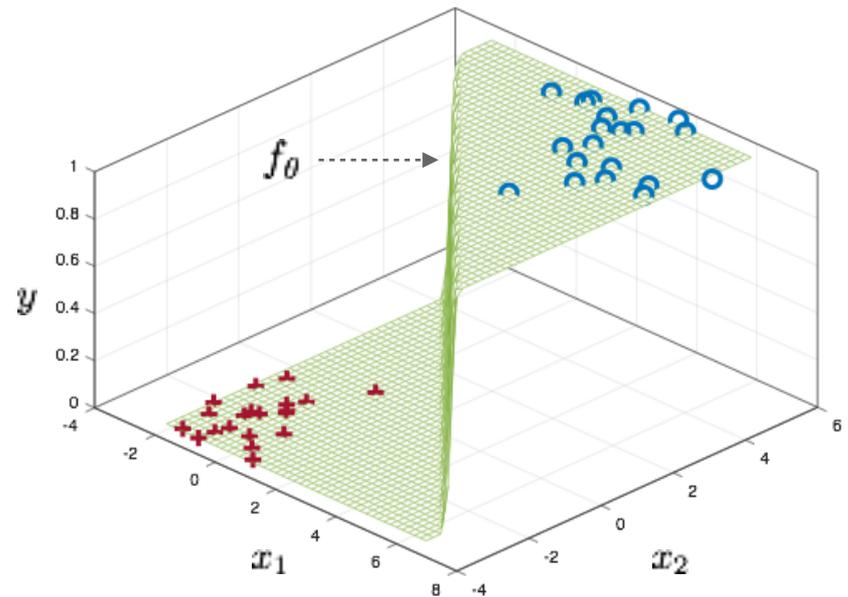
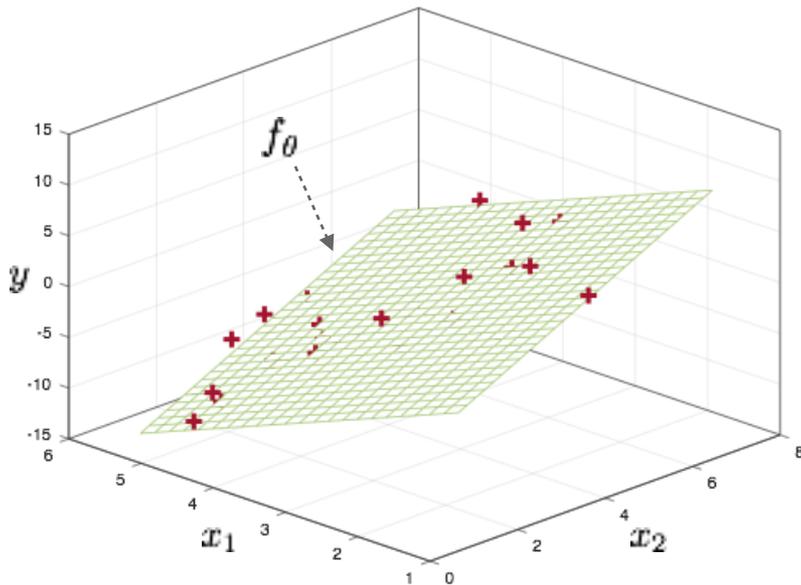
- *Le but est d'apprendre le modèle de prediction à partir des données*
- *Cela revient à trouver les “bonnes valeurs” des paramètres θ*



Regression ou classification

■ Types d'apprentissage supervisé

- **Regression** → Apprendre à prédire une réponse **continue**
- **Classification** → Apprendre à prédire une réponse **discrete**



Quiz

- Vous souhaitez développer des méthodes d'apprentissage pour résoudre les problèmes suivants.
 1. *Vous avez un large inventaire d'articles identiques. Vous voulez prédire combien de ces articles se vendront au cours des 3 prochains mois.*
 2. *Vous devez examiner les comptes de vos clients et décider pour chacun d'entre eux s'ils ont été piratés ou compromis.*
- Sont ils des problèmes de regression ou classification ?
 - A. *Problème 1 → classification - Problème 2 → classification*
 - B. *Problème 1 → classification - Problème 2 → regression*
 - C. *Problème 1 → regression - Problème 2 → classification*
 - D. *Problème 1 → regression - Problème 2 → regression*

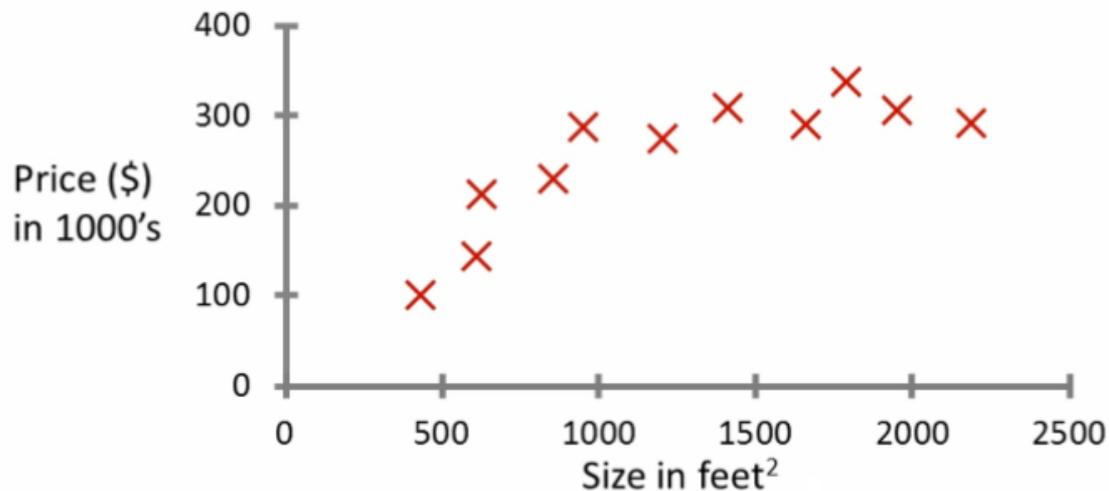
Panorama des méthodes d'apprentissage



Le problème de l'over-fitting

Choix du modèle (1/5)

- **PROBLEME** → Un ami a une maison de 750 feet²
 - *Quel est le prix qu'il pourrait obtenir pour la vente de sa maison...*
 - *... en connaissant le prix de quelques maisons dans le quartier ?*

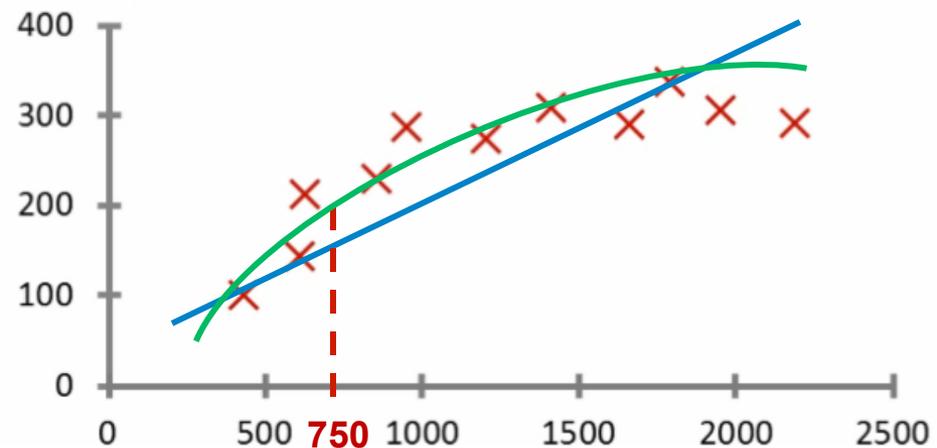


Choix du modèle (2/5)

- **PROBLEME** → Un ami a une maison de 750 m²
 - *Quel est le prix qu'il pourrait obtenir pour la vente de sa maison ?*

- **Approches possibles**

- *Modèle linéaire*
 - *la réponse est \$150'000*
- *Modèle quadratique*
 - *La réponse est \$200'000*

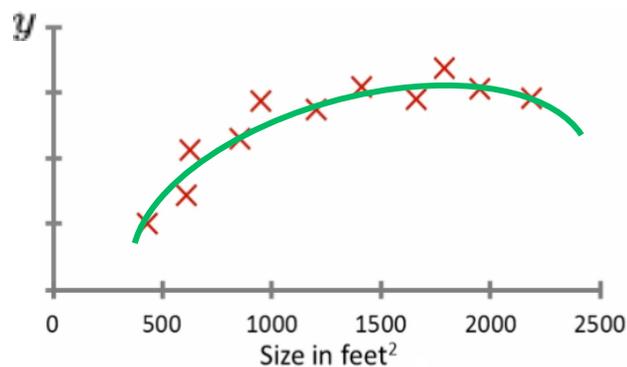


- Chacune est une façon d'effectuer une régression
 - *On y reviendra dans un autre cours...*

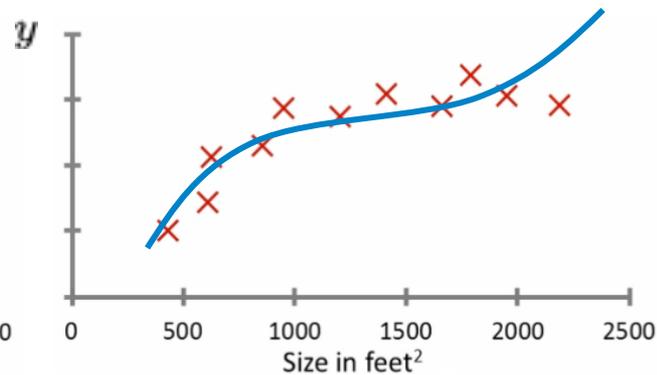
Choix du modèle (3/5)

- A priori, on peut choisir entre différents modèles
 - *quadratic* → le prix croit et puis décroît
 - *cubic* → le prix croit rapidement
 - *square root* → le prix croit lentement

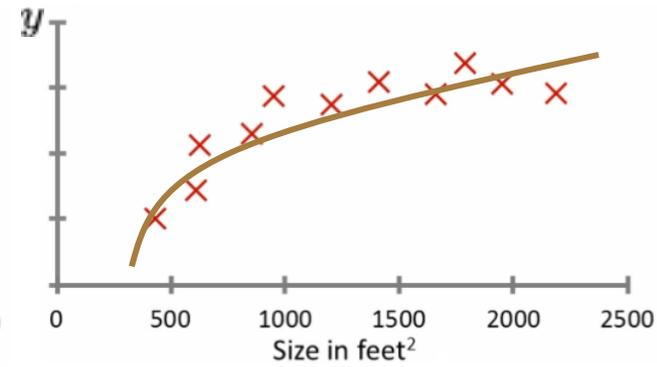
$$f_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



$$f_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$



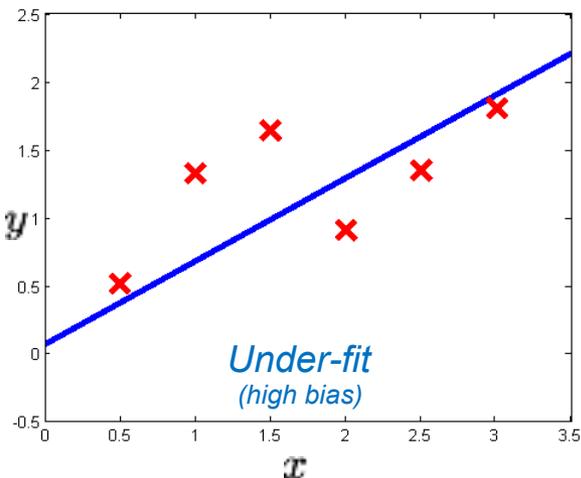
$$f_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 \sqrt{x}$$



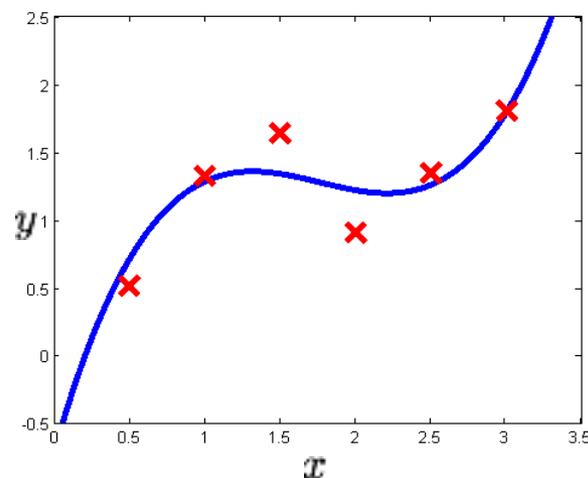
Choix du modèle (4/5)

- Quel est le modèle le plus approprié ?
 - *Under-fitting* → la prédiction est trop loin des vraies réponses
 - *Over-fitting* → la prédiction est trop proche aux vraies réponses

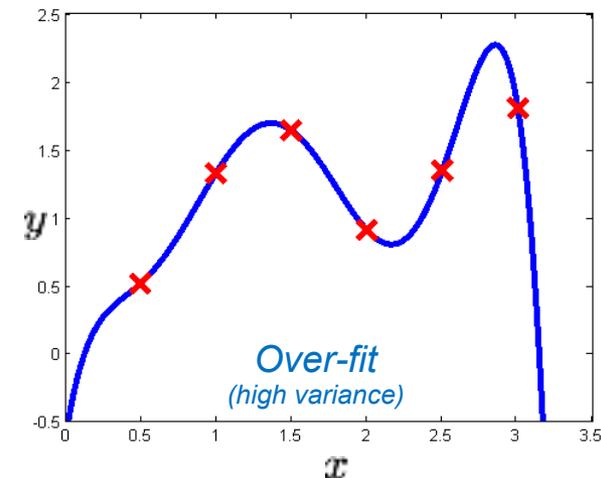
$$f_{\theta}(x) = \theta_0 + \theta_1 x$$



$$f_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_3 x^3$$



$$f_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_5 x^5$$



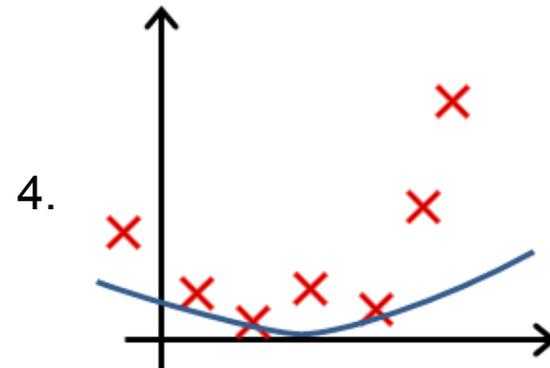
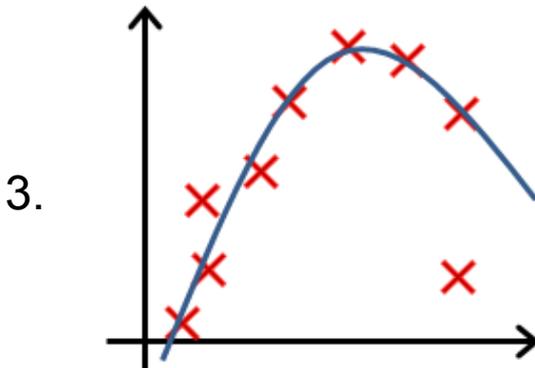
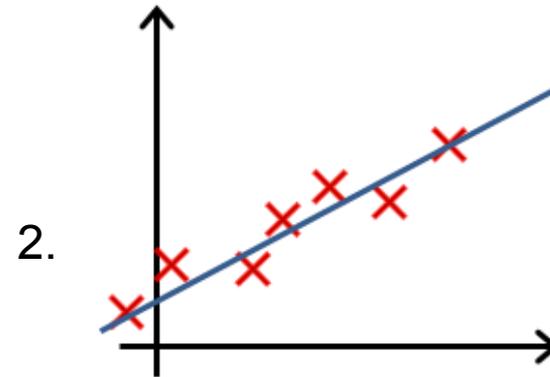
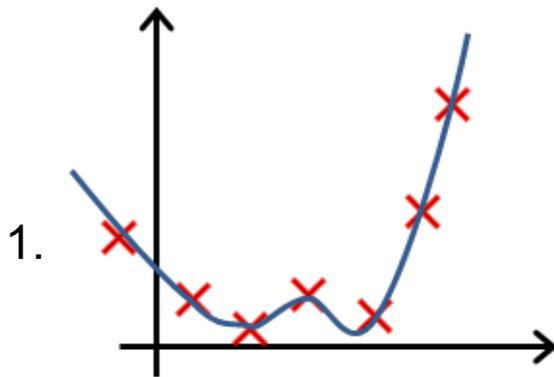
Choix du modèle (5/5)

- L'over-fitting est **mauvais** et il doit **toujours** être évité
 - *Le modèle prédit très bien les données d'apprentissage...*
 - *... mais il n'arrive pas à bien prédire les données de test*

- L'over-fitting est causé par les facteurs suivants
 - *Le modèle est très complexe (e.g., fortement non-linéaire)*
 - *Le nombre de paramètres du modèle est très élevé*
 - *La base d'apprentissage est trop petite*

Quiz

- Dans quelle figure le modèle souffre d'over-fitting ou d'under-fitting ?

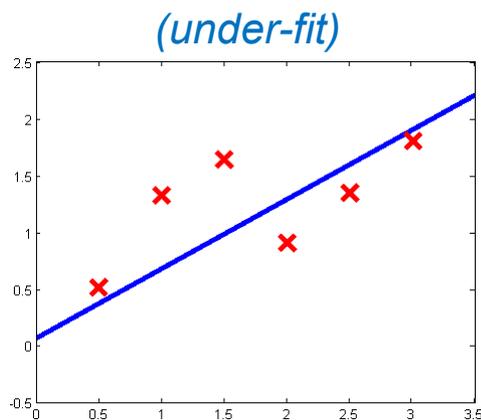


Evaluation des performances

Validation (1 / 3)

- Comment peut-on évaluer un modèle de prédiction ?
 - **WARNING** → On ne peut **pas** utiliser les données d'apprentissage !!!

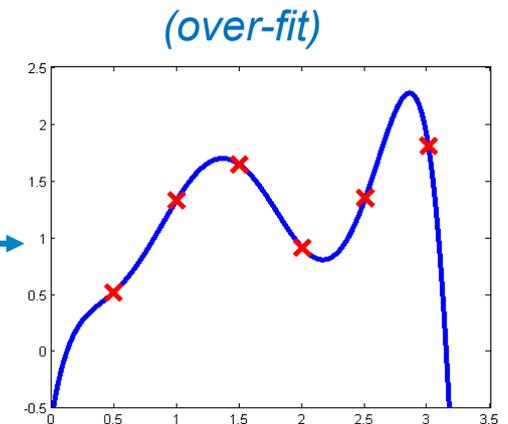
$$J_{\text{train}}(\hat{\theta}) = \frac{1}{N} \sum_{n=1}^N C(f_{\hat{\theta}}(\mathbf{x}^{(n)}), y^{(n)})$$



high

J_{Train}

low



Validation (2/3)

- L'approche correcte est d'utiliser des nouvelles données
 - *Training set* → Données destinées à l'apprentissage
 - *Validation set* → Données destinées au choix du "meilleur" modèle
 - *Test set* → Données destinées à l'évaluation du modèle retenu

Size	Price
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199

Diagram illustrating the data split for training, validation, and test sets:

- 60% → **training set**
- 20% → **validation set**
- 20% → **test set**

$$J_{\text{train}}(\hat{\theta}) = \sum_{n=1}^{N_1} C(f_{\hat{\theta}}(\mathbf{x}^{(n)}), y^{(n)})$$

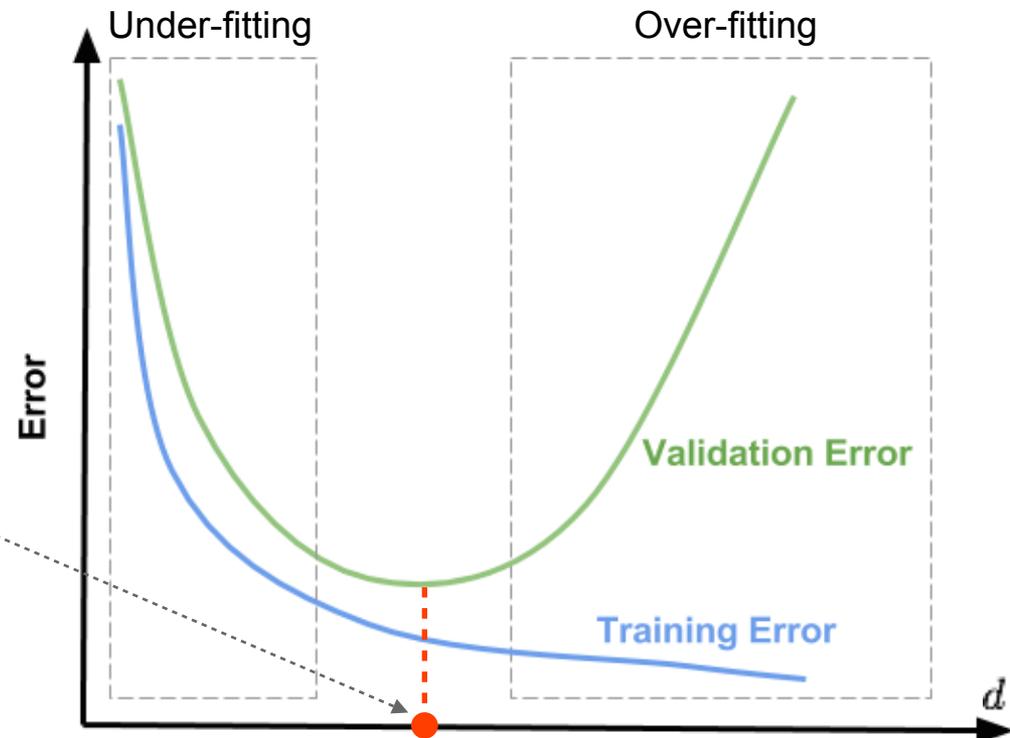
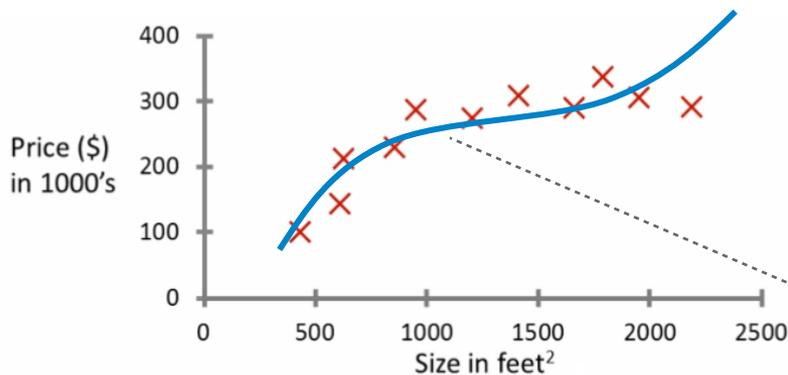
$$J_{\text{valid}}(\hat{\theta}) = \sum_{n=1}^{N_2} C(f_{\hat{\theta}}(\mathbf{x}_v^{(n)}), y_v^{(n)})$$

$$J_{\text{test}}(\hat{\theta}) = \sum_{n=1}^{N_3} C(f_{\hat{\theta}}(\mathbf{x}_t^{(n)}), y_t^{(n)})$$

Validation (3/3)

- **Example** → Régression polynomiale
 - *Quel degré du polynôme faut-il choisir ?*

$$f_{\theta}(x) = g(\theta_0 + \theta_1 x + \dots + \theta_d x^d)$$



Quiz

- Supposons que vous ayez appris un modèle de prediction qui marche très bien sur les données d'apprentissage.
- Que devrions-nous nous attendre ?
 - 1) *L'erreur sur les données d'apprentissage est **faible**, tandis que l'erreur sur les données de test est **élevé**.*
 - 2) *L'erreur sur les données d'apprentissage est **faible**, tandis que l'erreur sur les données de test est **faible**.*
 - 3) *L'erreur sur les données d'apprentissage est **élevé**, tandis que l'erreur sur les données de test est **faible**.*
 - 4) *L'erreur sur les données d'apprentissage est **élevé**, tandis que l'erreur sur les données de test est **élevé**.*

Systeme d'apprentissage

