

# PRÉDICTION DES PRIX IMMOBILIERS



King County  
USA



Khadija Ajimi, Le 21/12/2018

# SOMMAIRE



**Équipe**



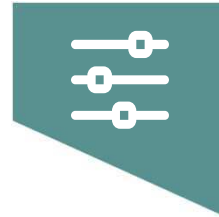
**Projet**



**Architecture  
Big Data**



**Intelligence  
Artificielle**



**Résultats**



**Conclusions &  
Perspectives**



# A PROPOS DE NOUS



**Djamel Gharbi** : Data architecte, Ingénieur Génie Civil

**Mohammed Ghiles** : Data ingénieur, Chef d'entreprise

**Anastasia Novikova** : Data scientist, PhD

**Khadija Ajimi** : Data scientist, Biologiste



# Le Projet : Estimation des Prix Immobiliers



1. Contexte
2. Motivations



# CONTEXTE ET MOTIVATION



- Projet Fil Rouge CBDDATA 4
- Dataset : <https://www.kaggle.com/harlfoxem/housesalesprediction/data>
- L'objectif de cette analyse est de prédire les prix des maisons dans ce comté.
- Le client est une entreprise de bâtiment du comté de King qui cherche à acheter des propriétés et à les revendre.
- Elle utilisera une application pour trouver des maisons moins chères à acheter.



# CONTEXTE ET MOTIVATION



Plus de 21000 de  
biens recensés



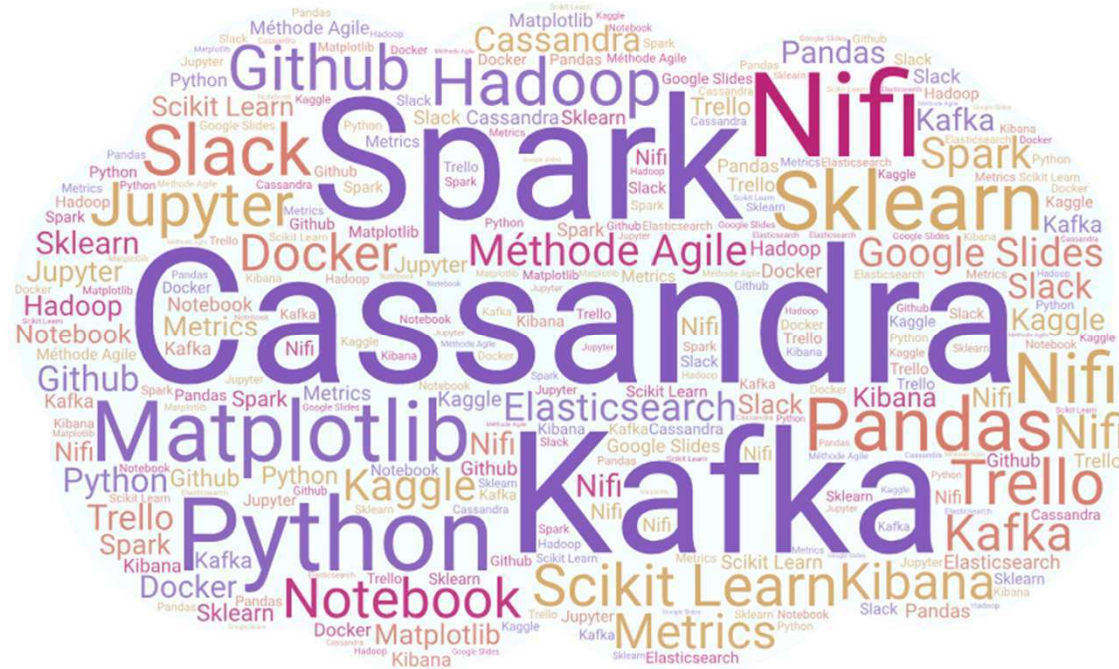
21 Paramètres



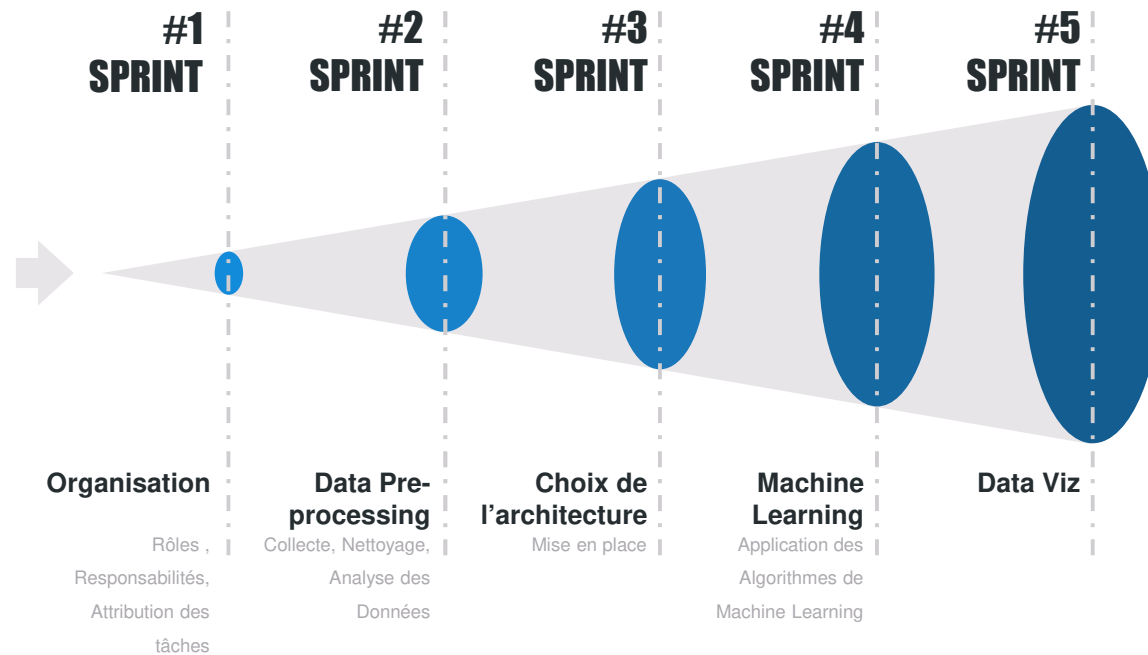
Transactions  
immobilières sur 2014-  
2015



Projet



# GESTION DE PROJET - PILOTAGE





# Architecture Big Data

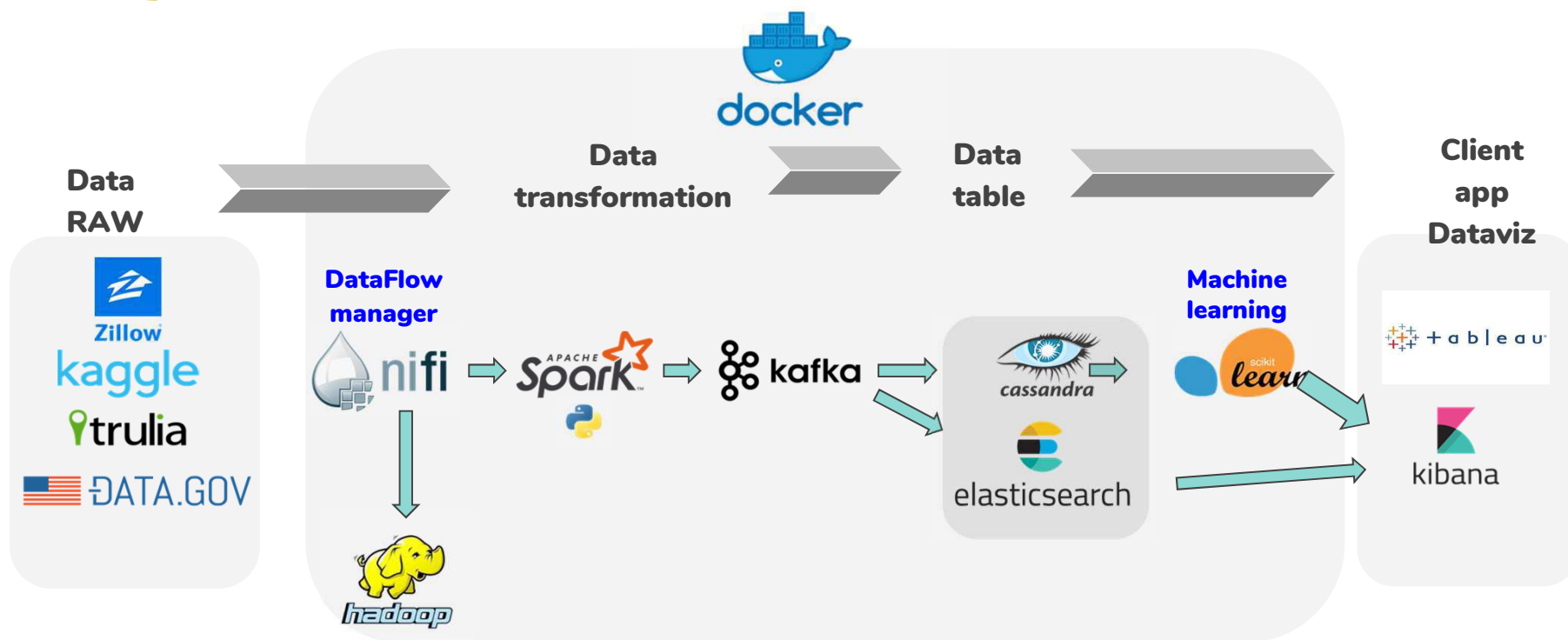


1. Schéma fonctionnel
2. Choix des outils Big Data



# 1.SCHEMA FONCTIONNEL

1  
1



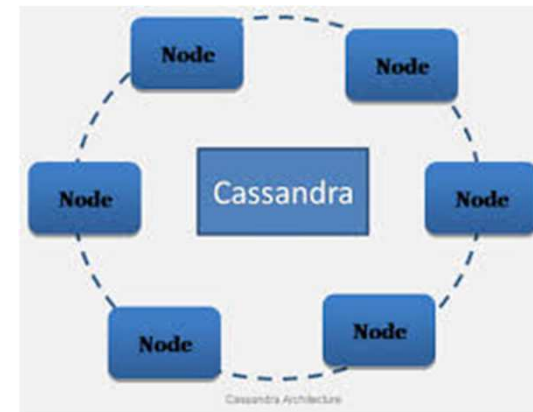
## 2. CHOIX DES OUTILS BIG DATA



Container pour la virtualisation des applications



Base de donnée NoSQL  
Haute disponibilité  
Scalabilité +++



# Intelligence Artificielle



1. Analyse exploratoire
2. Modèles de machine learning
3. Enrichissement du dataset





# ANALYSE EXPLORATOIRE

- Données manquantes: **Non**
- Type des données : **int/float**
- Transformer les variables : **Oui**
- Recherche des corrélations entre les variables **Oui**



# ANALYSE EXPLORATOIRE

kc\_house\_data.csv



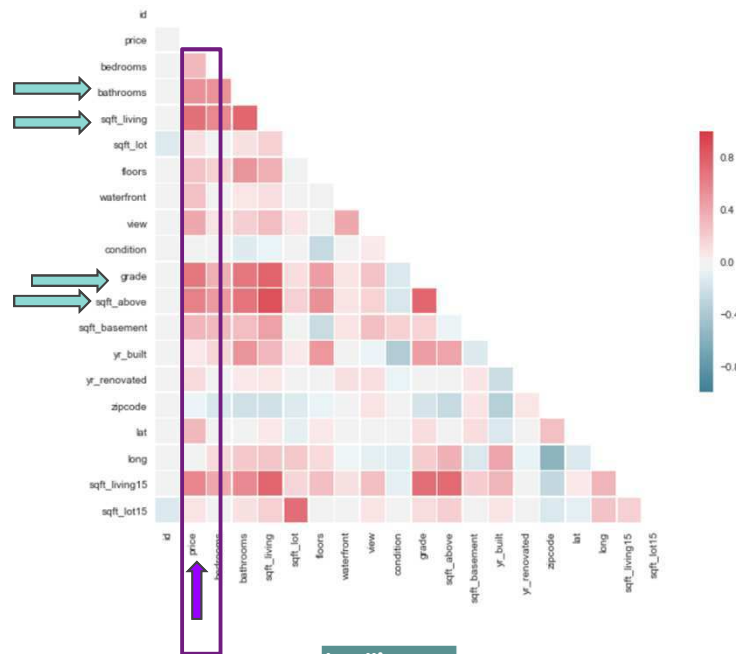
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0

21613 entrées  
21 variables





# MATRICE DE CORRELATION



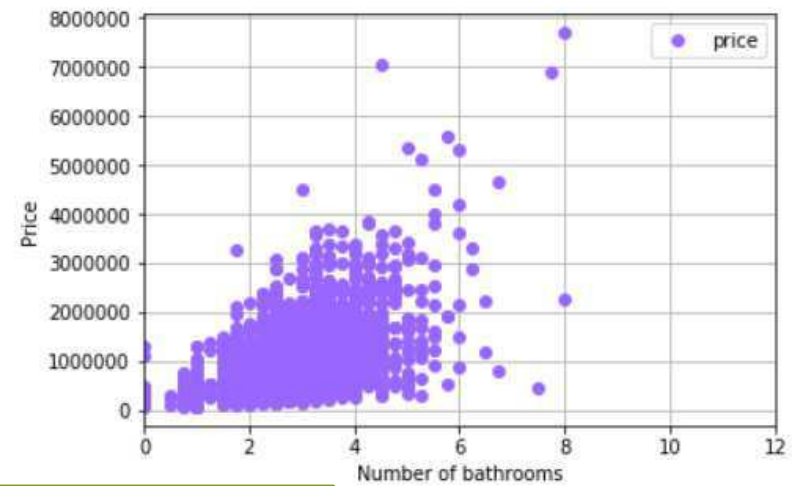
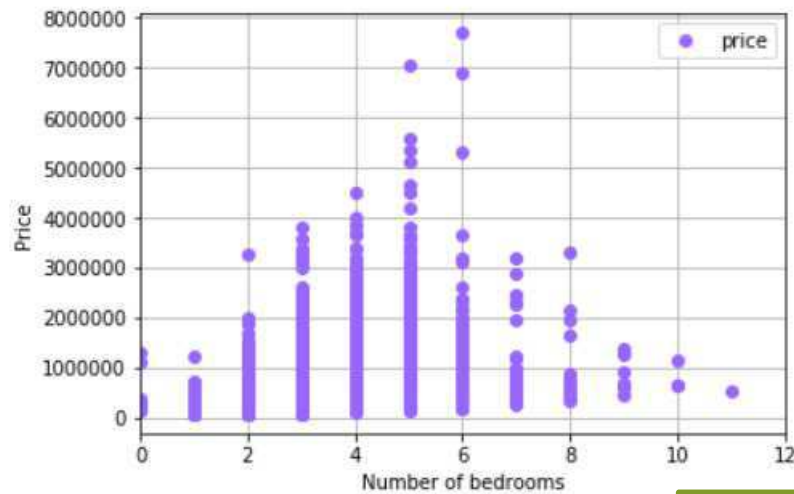
4 variables significatives :

- Salle de bain
- Superficie de la maison
- Superficie en dehors du sous-sol
- Note globale selon classement





## Prix vs Bedrooms, vs Bathrooms...



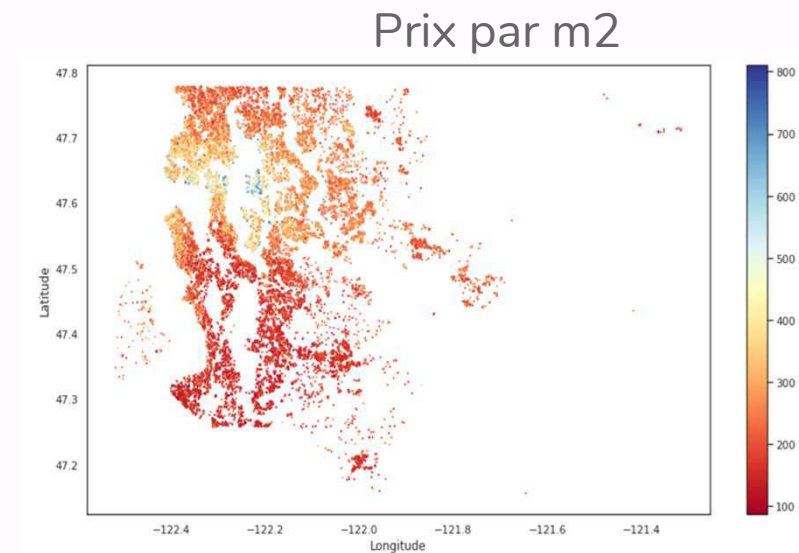
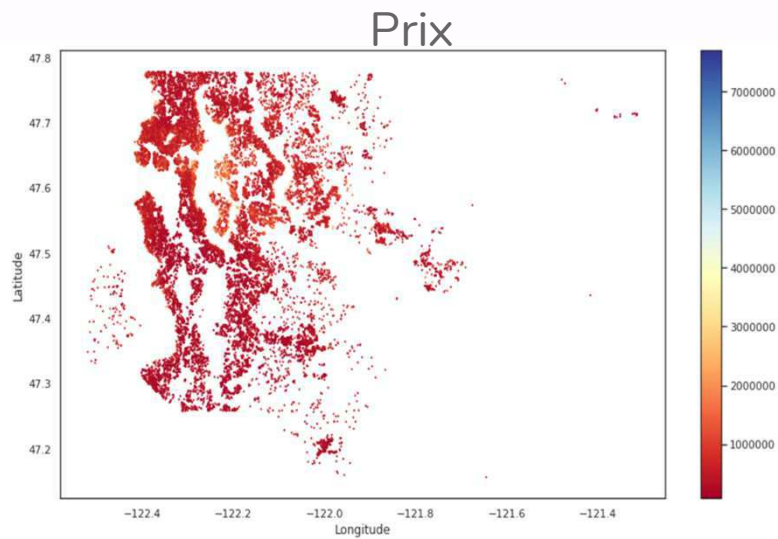
Nombre de chambres : 6  
Nombre de salles de bain : 6







# LE PRIX EN FONCTION DE LA SITUATION GEOGRAPHIQUE



Proximité : Lac Washington et Lac Sammamish





# MODELES DE MACHINE LEARNING

- Régressions Linéaires
- Arbre de décision, Random Forest
- Gradient Boosting



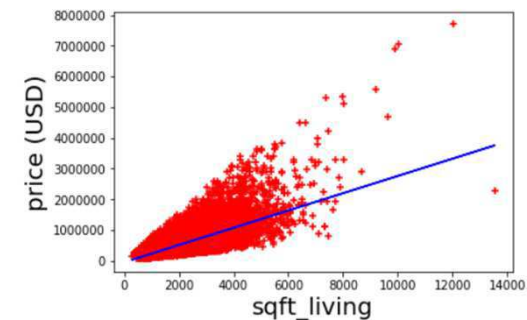
# REGRESSION LINEAIRE



- **variable d'entrée x** : surface
- **variable cible y** : prix



Relation linéaire simple reliant **y** à **x**



Même approche pour plusieurs variables :  
**Régression multivariée**

$$Y = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

$$Y = ax + b$$

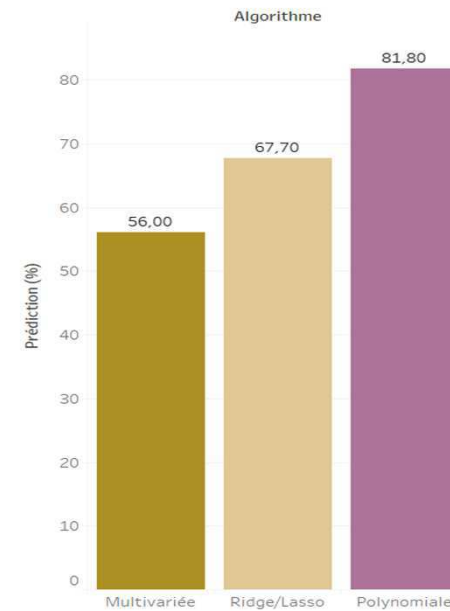


# REGRESSION LINEAIRE : RESULTATS



Algorithme	Prédiction
Multivariée	56,0%
Ridge / Lasso	67,7%
Polynomiale	81,8%

**Polynomiale > Ridge / Lasso > Multivariée**



# ARBRE DE DECISION

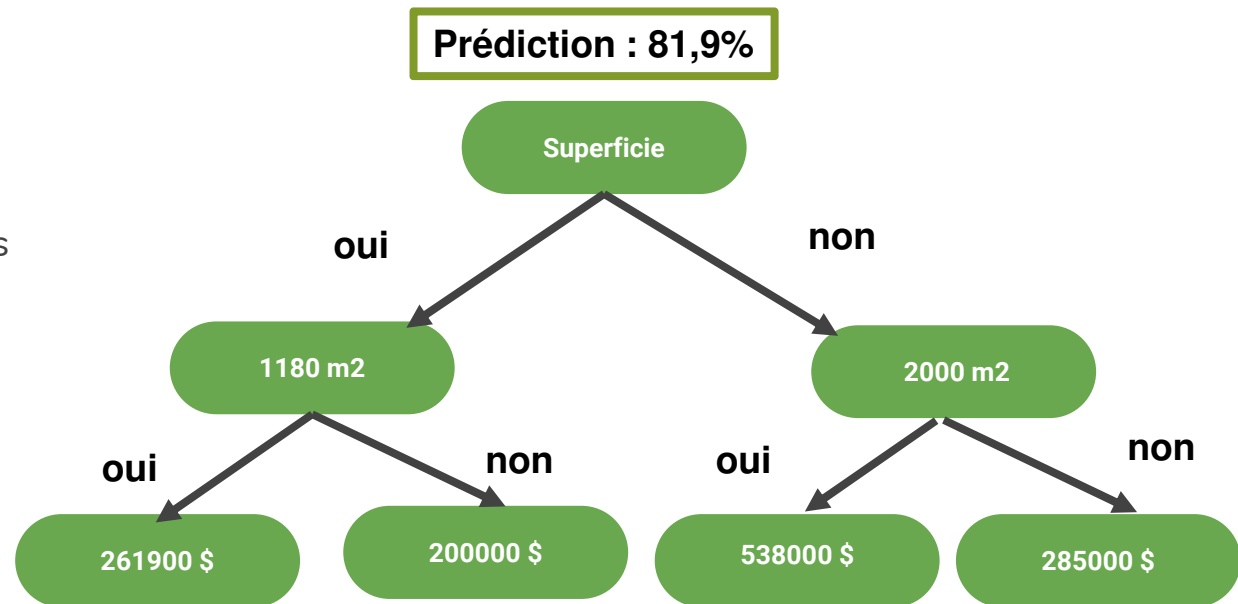


## Avantages :

- Peu de préparation de données
- Logique oui/non
- Performant sur de grands jeux de données

## Inconvénients :

- Arbre de décision très complexes
- Overfitting

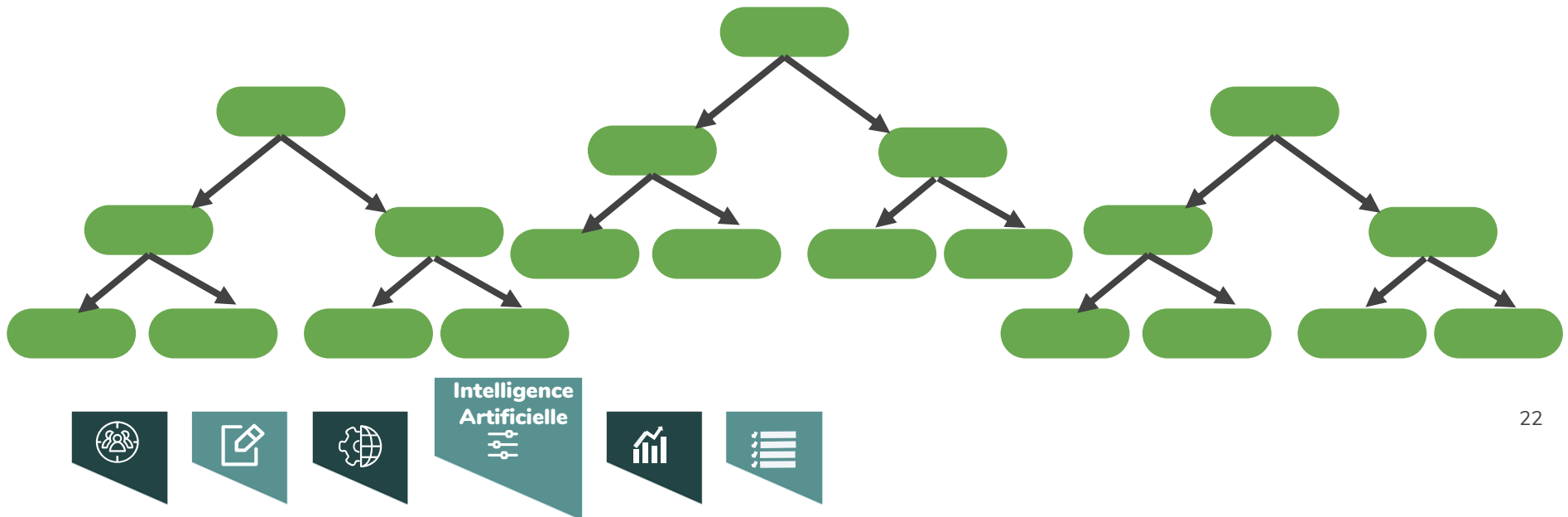


# RANDOM FOREST



Prédiction : 87,6%

- Algorithme **d'apprentissage supervisé**
- Combinaison de plusieurs arbres de décision : "Forêt aléatoire"



# RESULTATS : ARBRE DE DECISION / RANDOM FOREST

**Prédiction :**

Arbre de décision	Random Forest
81.9%	87.6 %

**Avantages du Random Forest :**

- Très stable : puissance de la “foule” car plusieurs arbres
- Fonctionne bien même avec les données manquantes (variables “view”, “waterfront”)

**Inconvénients du Random Forest :**

- Ressources de calcul ++++
- Temps de calcul ++++



# GRADIENT BOOSTING :



## Objectif :

Optimisation des arbres de décision

**Prédiction : 89%**

## Avantages Gradient Boosting :

- Permet de corriger les observations mal ajustées ou mal prédites par le Random Forest
- Mesure de l'erreur d'ajustement





# ENRICHISSEMENT DES DONNEES

1. Création de nouvelles variables à partir des variables existantes
2. Ajout des nouvelles lignes correspondant aux biens (scrapping)
3. Ajout des nouvelles variables à partir de Open Data : impôts, commerces, écoles de proximité, taux de criminalité (scrapping)

Dataset initial


Open Data




Application +++



# Résultats

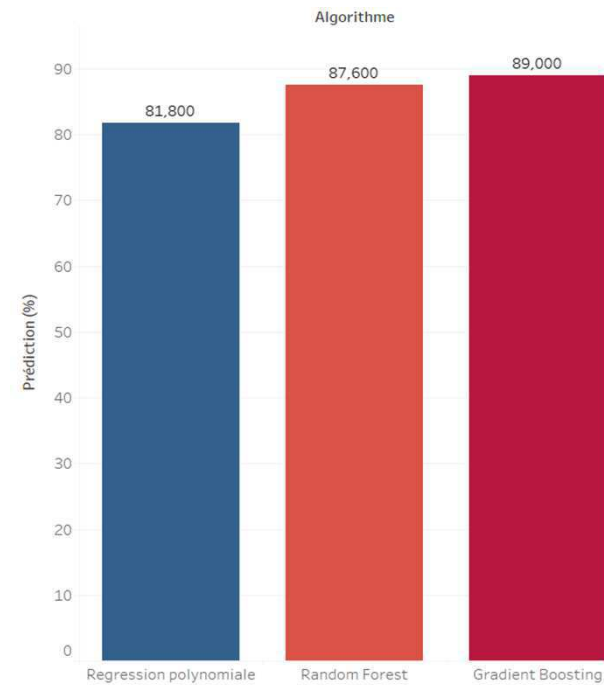


King County  
USA



# RESULTATS FINAUX

Algorithme	Prédiction
Régression polynomiale	81,8
Random Forest	87,6%
Gradient boosting	89,0%



# Conclusion & Perspectives



King County  
USA



## CONCLUSION ET PERSPECTIVES



1. Le meilleur modèle utilisé pour mettre en place cette application est celui de Gradient Boosting avec 89% de prédiction
2. Application sur Django  
<https://shrouded-scrubland-74851.herokuapp.com/homepage/>
3. Enrichissement du dataset avec deS données de transports, écoles, commerces et taux de criminalité



FichierÉditionAffichageHistoriqueMarque-pagesOutils?

test

←→↻🏠

https://shrouded-scrubland-74851.herokuapp.com/homepage/

Rechercher

Amazon Web Services...Catalogue en ligne - li...Sparkhttps://www.youtube.c...Nos cours - OpenClass...SlackIMTech Données envir...Formation Data Scient...Python docStatistiques et probab...notepad.pw / cbdata4...

Housing prices predictiondashboardNotebook

bedrooms\*

3

sqft\_living\*

1180

sqft\_lot\*

5650

latitude\*


47,5112

longitude\*

-122,257

Predict the Price

Predicted Price : \$352,266.44



**MERCI !**



King County  
USA

