

[Accueil](#) > [Cours](#) > [Initiez-vous au Machine Learning](#) > Plongez-vous dans la peau d'un Data scientist

## Initiez-vous au Machine Learning

10 heures



Moyenne

Mis à jour le 05/01/2022



## Plongez-vous dans la peau d'un Data scientist



01:37



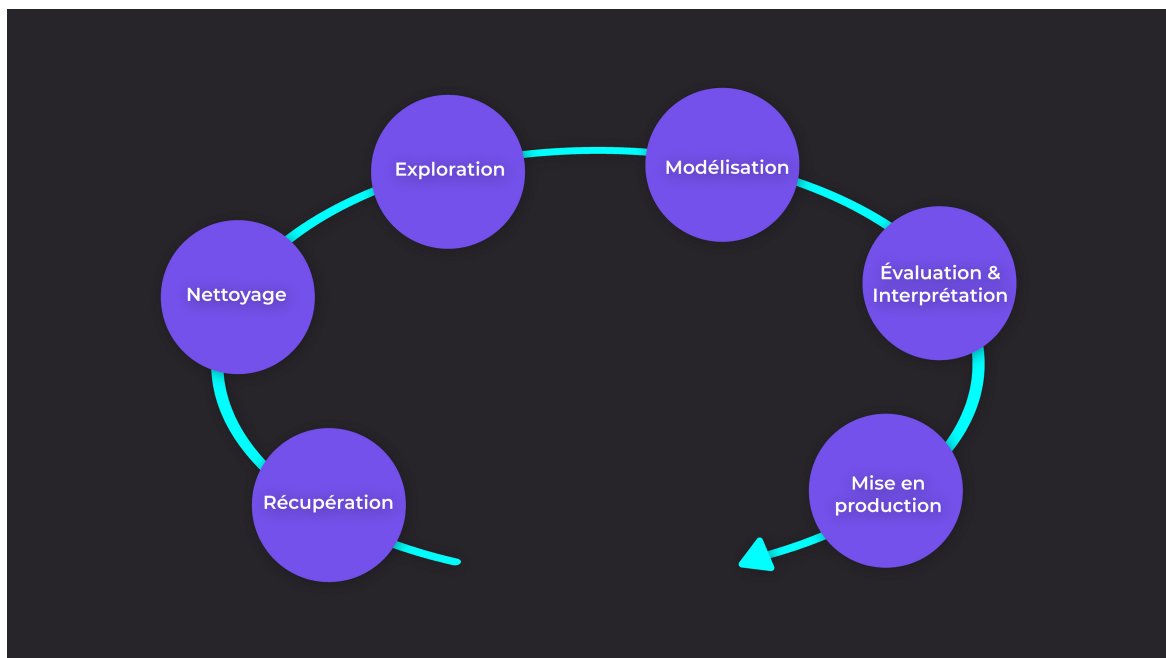
Le machine learning ne désigne en réalité qu'une partie du travail d'un data scientist. C'est pourquoi avant de rentrer dans le vif du sujet et de ne parler que de la partie machine learning, je vous propose de **faire un tour rapide du métier de data scientist**, afin de se situer.

Dans ce chapitre, nous allons prendre un peu de hauteur et observer en quoi consiste le **cycle habituel de travail des data scientists**, pour comprendre à quelle étape intervient le machine learning. C'est parti !

### Appréhendez le cycle de travail du data scientist



Le cycle de travail du data scientist peut se résumer par le schéma ci-dessous. Pour faire simple, on part de la réalité, on récupère les données, on les nettoie, on les explore, puis on utilise nos algorithmes pour créer de l'intelligence (artificielle) qui aide à la décision. Dans la suite, nous allons détailler ces différentes étapes et voir quels sont les différents métiers sur la chaîne de traitement de la donnée.



Cycle de travail du data scientist

## Récupérez les données



Une fois que vous êtes décidé à attaquer un problème, la première chose à faire est d'explorer toutes les pistes possibles pour récupérer les données. En effet, les données constituent l'expérience, les exemples que vous allez fournir à votre algorithme de machine learning, afin qu'il puisse apprendre et devenir plus performant.

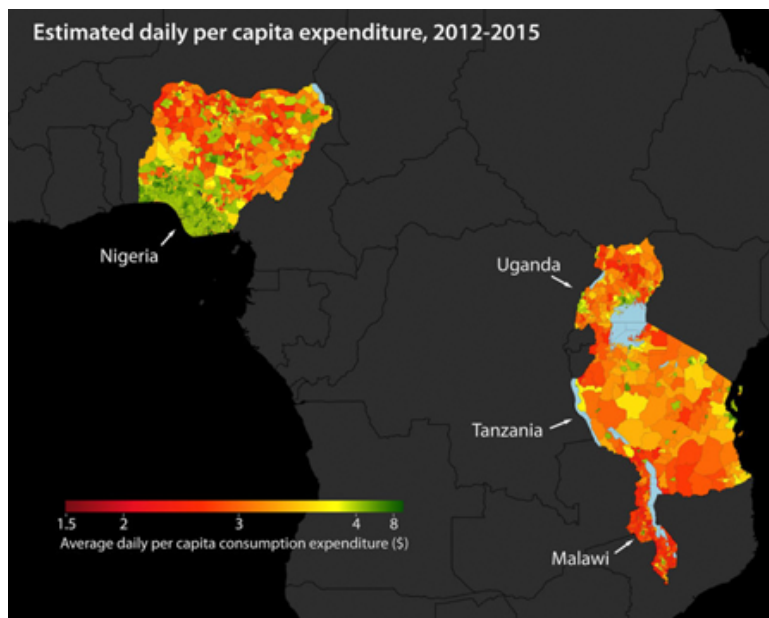
Dans la suite du cours, j'appellerai les données étudiées, destinées à alimenter un algorithme de machine learning, indifféremment **dataset** ou **jeu de données**.

**Tout doit passer au crible !** Les bases de données existantes, des données brutes alternatives (image, son, etc.), et même la création de nouveaux canaux d'acquisition de données. Essayez de trouver l'ensemble des variables qui impactent de près ou de loin le phénomène qui vous intéresse.

Vous trouverez ci-dessous quelques exemples, où les data scientists ont redoublé d'ingéniosité pour récupérer et utiliser leurs données de manière originale.

### Les images satellites pour évaluer le niveau de pauvreté

Des chercheurs ont utilisé le machine learning pour pouvoir cartographier les zones de pauvreté de manière automatique, simplement à partir [d'images satellites](#) !



Une cartographie de l'estimation de la consommation moyenne quotidienne (crédits : Neal Jean et al.)

## Les CAPTCHAs pour la digitalisation automatique de livres

Luis von Ahn, entrepreneur et chercheur, a créé un célèbre système de reCAPTCHA qui permettait à la fois aux sites web de valider que les formulaires étaient bien remplis par des humains, et qui alimentait en même temps la base de données d'un algorithme de digitalisation de livres. Grâce aux nombreux exemples renseignés directement par des humains, l'algorithme a fini par avoir suffisamment de données d'exemples pour réussir ensuite seul à retranscrire en texte des images scannées de livres, avec un taux d'erreur très faible.

Pour en savoir plus, vous pouvez consulter [cet article de 2018](#) sur la création du système de reCAPTCHA (article en anglais).

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

morning

Type the two words:

morning overtook

reCAPTCHA™  
stop spam.  
read books.

Exemple de reCAPTCHA

## Détectez l'illettrisme par l'utilisation du smartphone

Un chercheur norvégien a utilisé plusieurs types de données mobiles (tels que les SMS, le nombre de contacts, etc.) pour détecter les personnes illettrées dans les pays en voie de développement.

Pour en savoir plus, vous pouvez consulter [cet article de 2016 du site MIT Technology Review](#).

## Croisez les différentes sources de données

Dans beaucoup de cas, **l'innovation** en data science dans une entreprise vient de **l'originalité de l'utilisation des données** et du **croisement de différentes sources de données**. Pour cela, il faut dans l'idéal posséder une politique de gestion des données dans son entreprise la plus transparente possible. Pour les données, c'est comme pour les ressources humaines : les différents départements organisés en silos communiquent moins et innovent moins par rapport à un environnement où la transversalité est favorisée. Alors essayez d'éviter les **data-silos** !

## Nettoyez les données



Une fois les données trouvées, il faut passer à l'étape de nettoyage. Pour ne rien vous cacher, ce n'est pas l'étape la plus agréable du travail, mais ça ne la rend pas moins indispensable.

Nettoyer les données, c'est s'assurer qu'elles sont **consistantes**, sans **valeurs aberrantes** ni **manquantes**.

Pour aller plus loin dans cette étape, consultez le cours [Décrivez et nettoyez votre jeu de données](#).

Une autre étape nécessaire, en général, est l'aggrégation de ces données dans un **data lake**. Nettoyer les données signifie donc qu'elles sont toutes sous le même format, accessibles au même endroit et au bon moment.

Lorsque ces questions deviennent complexes, il faut faire appel au **data architect** qui, lui, possède une maîtrise technique pour réaliser ces différentes tâches. Ces ingénieurs des Big Datas sont responsables de la création et de l'administration de tous les systèmes techniques qui vont permettre la bonne exploitation des données.

Si vous souhaitez en savoir plus sur cet aspect technique de la data science, OpenClassrooms propose un parcours de formation de [Data Architect](#).

L'important, c'est de bien préparer le terrain pour les étapes suivantes. Ces étapes seront grandement simplifiées si ce travail fastidieux est bien effectué en amont.

## Explorez les données



Les données bien propres peuvent maintenant commencer à être explorées. Cette étape vous permet de **mieux comprendre les différents comportements** et de **bien saisir le phénomène sous-jacent**.

C'est vraiment une étape à ne pas négliger. Les meilleurs data scientists ne sont pas ceux qui connaissent les algorithmes les plus complexes, mais **ceux qui ont une très bonne connaissance des données** et ont préparé le terrain avec soin en amont.

À la fin de l'exploration, vous devrez être en mesure de :

- Proposer plusieurs hypothèses sur les causes sous-jacentes à la génération du dataset : "suite à

l'exploration, il y a clairement une relation entre X et Y".

- Proposer plusieurs pistes de modélisation statistique des données, qui vont permettre de résoudre la problématique de départ considérée.
- Proposer si nécessaire de nouvelles sources de données qui aideraient à mieux comprendre le phénomène.

C'est dans les phases de nettoyage et d'exploration des données que les data scientists passent le plus clair de leur temps.

Lorsque l'on a simplement besoin de comprendre ses données et les explorer, on peut faire appel à un **data analyst**. Ou bien un data analyst peut effectuer des études préliminaires avant de laisser le travail de modélisation au data scientist. Si vous souhaitez vous former à ce métier, OpenClassrooms propose un parcours de formation de [Data Analyst](#).

## Modélisez les données à l'aide du machine learning



Nous pouvons enfin rentrer dans la partie la plus intéressante du métier, c'est-à-dire la création du modèle statistique associé aux données qui nous intéressent ! C'est ce qu'on appelle le **machine learning** (ou apprentissage automatique).

Mais ça veut dire quoi "modélisation statistique des données" ?

En machine learning, et en data science plus généralement, l'objectif est de trouver un modèle (stochastique ou déterministe) du phénomène **à l'origine** des données. C'est-à-dire qu'on considère que **chaque donnée observée est l'expression d'une variable aléatoire générée par une distribution de probabilité**.

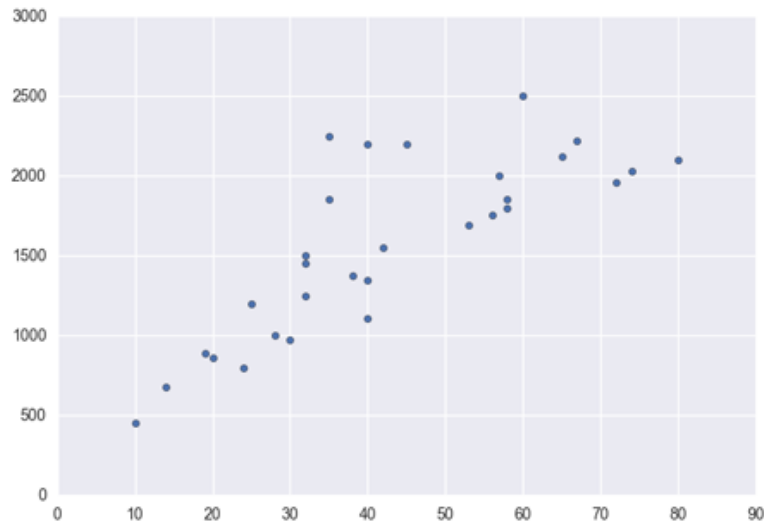
Si vous n'avez jamais entendu parler de **variable aléatoire** ou de **distribution de probabilité**, vous allez avoir des difficultés à suivre le reste du cours. Je vous conseille de suivre en amont un cours d'introduction aux probabilités et aux statistiques. Vous en trouverez dans le parcours [Data Analyst](#) sur OpenClassrooms.

Le mieux pour expliquer ce que ça signifie est de prendre un petit exemple simple. Imaginez que vous voulez savoir si vous payez trop cher votre loyer. Vous avez récupéré sur un site de location une trentaine de prix des locations disponibles, ainsi que la surface associée :

loyer mensuel (en €)	surface (en $m^2$ )
1500	32
2120	65
2500	60
...	...

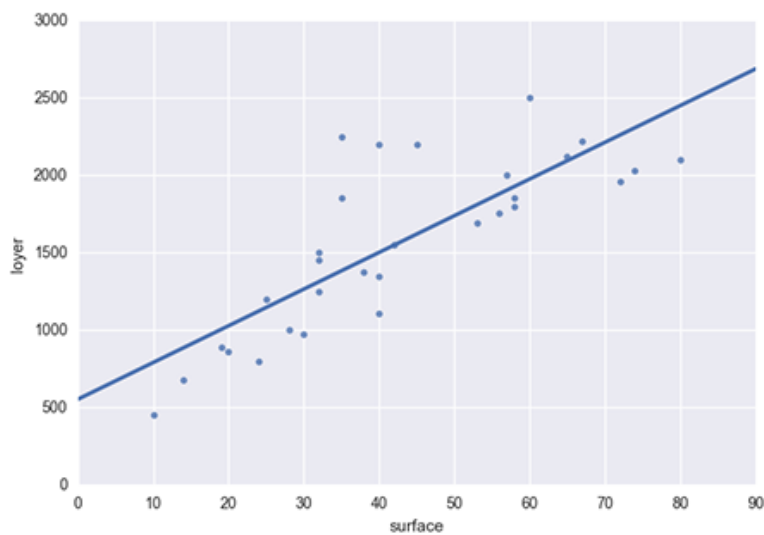
Bien sûr, en réalité d'autres paramètres seraient probablement à prendre en compte (parties communes, voisinage, évolution des loyers au cours du temps, etc). Le but est ici d'appréhender un modèle simplifié afin de comprendre rapidement ce que veut dire "modéliser un phénomène".

Si l'on affiche maintenant ces différents points sur un graphe qui représente le montant du loyer en fonction de la surface, on obtient le graphique suivant :



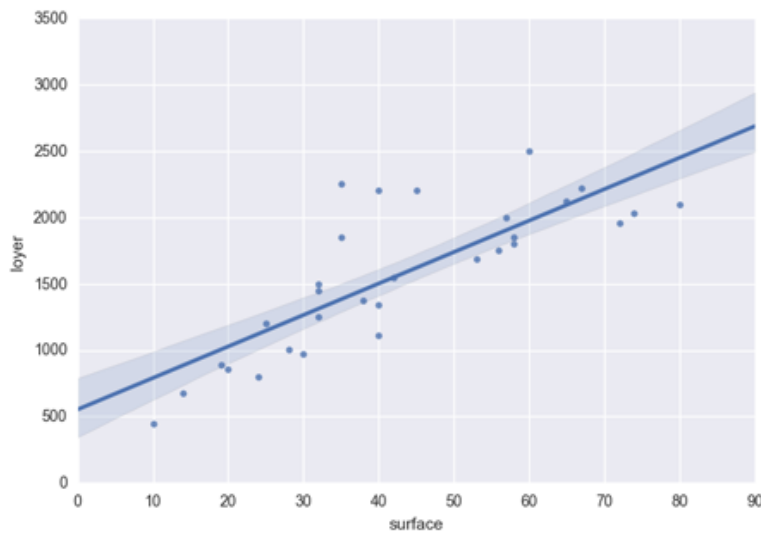
Loyer mensuel en fonction de la surface du logement

Comme on pouvait s'y attendre, on remarque une augmentation relativement **linéaire** du loyer par rapport à la surface de l'appartement. Une première **modélisation** simple du phénomène (le prix du loyer) serait donc simplement de considérer la droite la plus "proche" de l'ensemble des points.



La droite de régression correspondant à la modélisation du nuage de points

La droite représente donc notre **modèle** du phénomène, auquel nous pouvons ajouter l'intervalle de confiance dans lequel on pense que se trouve la droite.



L'intervalle de confiance (à 90 %)

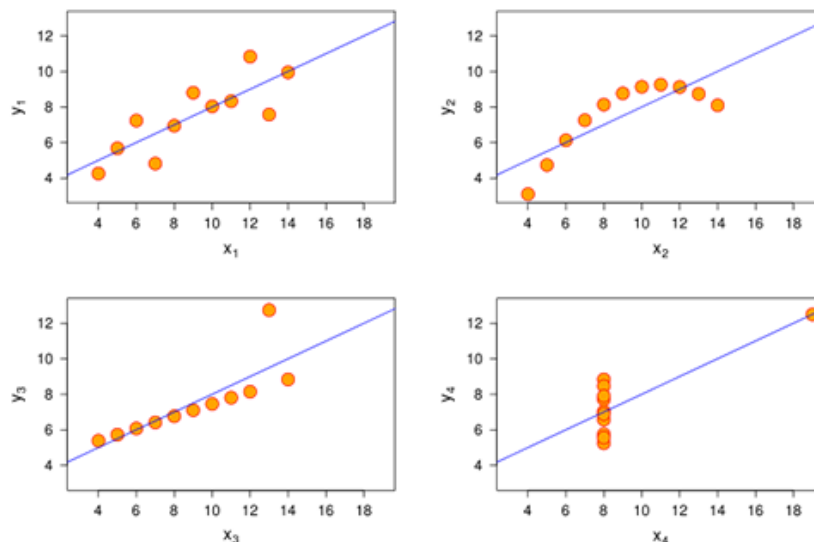
Pour résumer, le travail de modélisation consiste à trouver le bon modèle statistique (ici la droite et son intervalle de confiance) qui *colle le mieux aux données d'exemple*. Le machine learning en particulier intervient pour trouver ce modèle de manière *automatisée*.

## Évaluez et interprétez les résultats



Une fois un premier travail de modélisation effectué, la suite de l'étude s'effectue par **l'évaluation de la qualité de notre modèle**, c'est-à-dire sa capacité à représenter avec exactitude notre phénomène, ou a minima sa capacité à résoudre notre problématique.

Une représentation connue qui souligne la nécessité de l'évaluation est le **quartet d'Anscombe**. Il permet de montrer visuellement que pour 4 jeux de données très différents, on obtient la même droite de régression.



Le quartet d'Anscombe illustre bien le fait que si l'on n'examine pas assez les données, et qu'on ne mesure pas de la bonne manière l'erreur de son modèle, on peut facilement arriver à des aberrations de modélisation.

Il y a parfois clairement un problème dans notre modèle, qui ne capture pas l'essence du phénomène. Pour nous aider à évaluer les résultats, mesurer l'erreur de notre modélisation vis-à-vis de nos données d'exemple constitue un premier indicateur de qualité. Dans les cas ci-dessus, il faudrait clairement changer le modèle

d'une droite que nous avons décidé au départ !

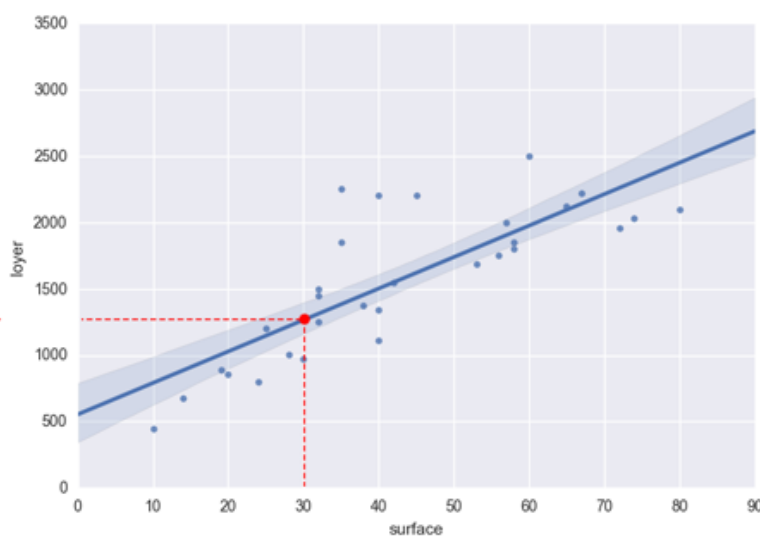
C'est donc un jeu d'allers-retours entre modélisation et évaluation qui s'effectue pour obtenir les performances les plus satisfaisantes possibles. Il est même possible, dans certains cas, de remettre en question certaines hypothèses de départ et de repartir dans une phase d'exploration pour mieux comprendre les données.

## Déployez le modèle en production



Une fois qu'on est satisfait de la qualité des performances de notre modèle, on va pouvoir passer à l'étape suivante, qui est le rendu de nos résultats et le potentiel déploiement du modèle en production. Imaginez que vous trouvez que votre modèle d'évaluation des loyers est très performant et mériterait d'être partagé à plus de monde. Vous décidez donc de le déployer sur un serveur où tout le monde pourra obtenir une estimation de son loyer selon votre modèle, et ainsi déterminer s'il paie plus ou moins que les prix du marché ! Cela l'aidera sûrement dans sa décision de déménager. 🤖

Comment cela fonctionne-t-il en pratique ? C'est assez simple, il vous suffit de récupérer les paramètres de votre modèle et de faire passer la surface de l'appartement en entrée du modèle, afin d'obtenir le loyer associé en sortie, en suivant la droite.



Imaginez qu'un appartement a une surface de 30 mètres carrés (point en rouge), une estimation légitime du loyer se situerait aux alentours de 1300 euros selon notre modèle.

Pour des modèles plus complexes, le fonctionnement reste le même. Si vous voulez appliquer votre travail à de nouvelles données, il vous suffit de passer les nouvelles entrées dans votre modèle (qui est en principe un ensemble de transformation des valeurs d'entrées) afin d'obtenir une sortie.

Là encore, si ce passage en production est complexe, que ce soit en termes d'échelle, de contrainte de rapidité de calcul ou de sortie de résultats, il faut faire appel à un data architect qui sera responsable d'industrialiser le prototype que vous lui fournirez.

## Le professeur

### En résumé

Yannis Chaouche



La data science est un nouveau domaine de travail qui augmente les capacités d'analyse classique, afin d'aider les entreprises à prendre des décisions plus informées. Elle s'appuie pour cela sur des données utiles et ne peut s'appliquer que dans certaines problématiques précises qui gagnent à utiliser ce type de méthodes.



---

Au sein du cycle de travail du data scientist, le **machine learning** désigne l'ensemble des méthodes de **modélisation statistique à partir des données**.

OPENCLASSROOMS

Qui sommes-nous ?

J'AI TERMINÉ CE CHAPITRE ET JE PASSE AU SUIVANT

Alternance

Financements

← **DÉCOUVREZ LE DOMAINE DE LA DATA**

Expérience de formation

**SCIENCE**

**IDENTIFIEZ LES DIFFÉRENTES ÉTAPES DE**  
**MODÉLISATION** →

Forum

Blog [↗](#)

Presse [↗](#)

---

## OPPORTUNITÉS

Nous rejoindre [↗](#)

Devenir mentor [↗](#)

Devenir coach carrière [↗](#)

## AIDE



FAQ

---

## POUR LES ENTREPRISES

Formation, reconversion et alternance

## EN PLUS

Boutique [↗](#)

Mentions légales


Conditions générales d'utilisation

Politique de protection des données personnelles

Cookies

Accessibilité

---

 Français ▼

